Correlations in thermodynamics and evolution of proteins

Korrelationen in Thermodynamik und Evolution von Proteinen
Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation von Dipl.-Phys. Jonas Minning aus Bendorf am Rhein
Juni 2012 — Darmstadt — D 17

TECHNISCHE
UNIVERSITÄT
DARMSTADT

AG Porto/AG Drossel
Institut für Festkörperphysik
Fachbereich Physik

Correlations in thermodynamics and evolution of proteins
Korrelationen in Thermodynamik und Evolution von Proteinen

Genehmigte Dissertation von Dipl.-Phys. Jonas Minning aus Bendorf am Rhein

1. Gutachten: Dr. Markus Porto
2. Gutachten: Prof. Dr. Barbara Drossel

Tag der Einreichung: 15.05.2012
Tag der Prüfung: 04.06.2012

Darmstadt — D 17

Abstract

An important prerequisite for the biological function of a protein is the thermodynamic stability of its three-dimensional structure, the so-called native state. By adjusting the amino acid sequence the stability can be optimized by two different strategies. While positive design increases the stability with respect to unfolding by decreasing the free energy of the native state, negative design increases the free energy of misfolded structures in order to optimize the stability against misfolding. One stability can be optimized only at the expense of the other, thus optimal stability demands a trade-off between the two strategies.

In the first part of this work, negative design in naturally occurring proteins was investigated using a simple energy model based on contact interactions of amino acids. The calculation of the free energy of the misfolded ensemble is difficult due to the large number of misfolded structures. A widely used model to describe the free energy of the misfolded ensemble is the Random Energy Model (REM), which assumes contacts to be uncorrelated and to occur with equal frequency. This is, however, an inaccurate description, as the probability of contact decreases with increasing distance in the sequence and the formation of a contact in a misfolded structure is correlated with other contacts. The first part of the thesis investigates how contact frequency and contact correlation affect negative design. Here, the free energy of the misfolded ensemble is approximated by a cumulant expansion, where contact frequency and contact correlation are explicitly included. In addition, it is investigated how the description of optimal hydrophobicity profiles, which have maximal stability in the native state, can be enhanced by the inclusion of contact correlations. The detailed description of the misfolded ensemble can help to improve the design of sequences or allows a more accurate modeling of protein evolution.

Since protein sequences change during evolution, correlated substitutions of amino acids at different sites in the protein — in the literature often referred to as correlated mutations — give insight into the native structure and function of a protein. However, there was no theoretical description to quantify the effects of the physical constraints of structure and folding stability on correlated mutations in protein sequences. In the second part, a model is studied which quantitatively predicts the correlated mutations from constraints on the folding stability. The model is based on maximizing the sequence entropy, which is approximated by a cluster expansion up to second order. The model is tested using data from computer simulations and a statistical analysis of proteins from the Protein Data Bank. In particular, the determination of the model parameters allows an interpretation of the correlations in terms of both design strategies that characterize sequence evolution. The model can help to distinguish native from non-native contacts based on correlated mutations, thus improving the prediction of contacts and hence the prediction of protein structures. In addition, the model could be helpful to distinguish between correlated mutations that result from the folding stability or other selective pressures.

Zusammenfassung

Eine wichtige Voraussetzung für die biologische Funktion eines Proteins ist die thermodynamische Stabilität dessen dreidimensionaler Struktur, der sogenannte native Zustand. Durch Anpassung der Aminosäurensequenz lässt sich die Stabilität durch zwei verschiedene Prinzipien optimieren. Während positives Design die Stabilität gegen Entfaltung erhöht, indem es die freie Energie des nativen Zustandes erniedrigt, versucht negatives Design die freie Energie missgefalteter Strukturen zu erhöhen, um so die Stabilität gegen Missfaltung zu optimieren. Eine Stabilität kann nur auf Kosten der anderen optimiert werden, so dass ein Kompromiss zwischen beiden Prinzipien gefunden werden muss.

Im ersten Teil der Arbeit wurde anhand eines einfaches Energiemodells, das auf Kontaktwechselwirkungen von Aminosäuren beruht, negatives Design in natürlich vorkommenden Proteinen untersucht. Die Beschreibung der freie Energie des missgefalteten Ensembles ist auf Grund der großen Anzahl von missgefalteten Strukturen schwierig. Ein weit verbreitetes Modell zur Berechnung der freien Energie des missgefalteten Ensembles ist das Random Energy Model (REM), das die möglichen Kontakte zwischen Residuen als unkorreliert und mit gleicher Häufigkeit vorkommend annimmt. Dies ist jedoch nur eine ungenaue Beschreibung, vielmehr vermindert sich die Kontaktwahrscheinlichkeit mit wachsenden Abstand in der Sequenz und das Formen eines Kontakts in einer missgefalteten Struktur ist mit anderen Kontakten korreliert. Im ersten Teil der Arbeit wird untersucht, wie Kontakthäufigkeit und Kontaktkorrelationen negatives Design beeinflussen. Dabei wurde die freie Energie des missgefalteten Ensembles in einer Kumulantenentwicklung approximiert, in die explizit Kontakthäufigkeiten und Kontaktkorrelationen einbezogen werden. Zudem wird untersucht, inwiefern sich die Beschreibung optimaler Hydrophobitätsprofile, die maximale Stabilität der nativen Struktur erreichen, durch die Einbeziehung von Kontaktkorrelationen verbessern lässt. Die genauere Beschreibung des missgefalteten Ensembles kann zu einer Verbesserung von Design von Sequenzen oder einer genaueren Modellierung von Proteinevolution beitragen.

Da Proteinsequenzen sich in Laufe der Evolution verändern, liefern korrelierte Substitution von Aminosäuren an verschiedenen Plätzen im Protein – in der Literatur oft mit korrelierten Mutationen bezeichnet – Einsicht in die native Struktur und Funktionen eines Proteins. Jedoch gab es keine bislang noch keine theoretische Beschreibung, die die Auswirkungen der physikalischen Beschränkungen durch Struktur und Faltungsstabilität auf Korrelationen in Proteinsequenzen quantifizieren. In zweiten Teil der Arbeit wird ein Modell untersucht, das korrelierte Mutationen aus Bedingungen an die Faltungsstabilität quantitativ vorhersagt. Die Grundlage des Modells ist die Maximierung der Sequenzentropie, die durch eine Cluster-Enwicklung bis zur zweiten Ordnung approximiert wird. Das Modell wird anhand von Daten aus Computersimulationen und einer statistischen Analyse von Proteinen aus der Protein Data Bank getestet. Insbesondere erlaubt die Bestimmung der Modellparameter eine Interpretation der Korrelationen in Bezug auf die beiden Designstrategien, die die Sequenzevolution prägen. Das Modell kann Hinweise darauf liefern, wie man native von nicht-nativen Kontakten unterscheidet, und so zur Verbesserung der Vorhersage von Kontakten und damit von Proteinstrukturen beitragen. Zudem könnte das Modell dabei behilflich sein, zwischen korrelierten Mutationen zu unterscheiden, die aus der Faltungsstabilität oder von anderen Selektionsdrücken herrühren.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Proteins are molecular machines that perform a wide array of functions within a living organism. They build structural units of the cell, move whole organism in muscles, act as signal carriers in the cell cycle, and catalyze chemical reactions in the cell. Although the functions differ, all proteins are built by the same principle: They consist of a chain of amino acids, whose different composition gives rise to the large variety of proteins.

Critical to the function of the protein is the specific interaction with other molecules present in their environment, which is mediated by the three dimensional structure of a protein. Astonishingly, most proteins fold into a unique three dimensional structure, the so-called native state, which is solely determined by the sequence of amino acids [1]. The folding process is governed by physical laws. However, due to its complexity this process cannot be described theoretically and is still a topic of today's research. An interesting question is how the protein selects the native structure among the many different possible conformations of its chain. Indeed, the correct folding of the protein chain is of medical importance as the misfolding of proteins is linked to many diseases, most prominently Alzheimer's [2].

Even though it does not answer questions of the actual process of folding, equilibrium physics can give insight into the folding of proteins. The dominance of the native state in the ensemble of possible conformations requires the native state to be thermodynamically stable, that is, the free energy of the native state has to be lower than the free energy of misfolded conformations. However, the large number of conformations of the chain renders a detailed modeling of the misfolded ensemble difficult. In this thesis, it is investigated to which extent natural proteins are optimized with respect to misfolding. To this end, the free energy of the misfolded ensemble is described by a model that respects statistical properties of misfolded conformations in the first part of this thesis.

Proteins are changing during evolution, which allows them to adapt to a changing environment and to fulfill new functions. On the other hand, certain functions are required for the survival of the organism and the proteins function has to be conserved, which often means that the proteins structure is conserved. This selection pressure shapes the statistical properties of sequences that a protein adopts during evolution. Thus, the observation of these properties can give insight into functional and structural properties of proteins. For a deeper understanding it is important to have theoretical models that connect the evolutionary and the structural aspect. In the second part of this thesis, I investigate a model the coevolution of amino acids at different sites, i.e., positions, in the protein, which arise from the thermodynamic stability of the native state.

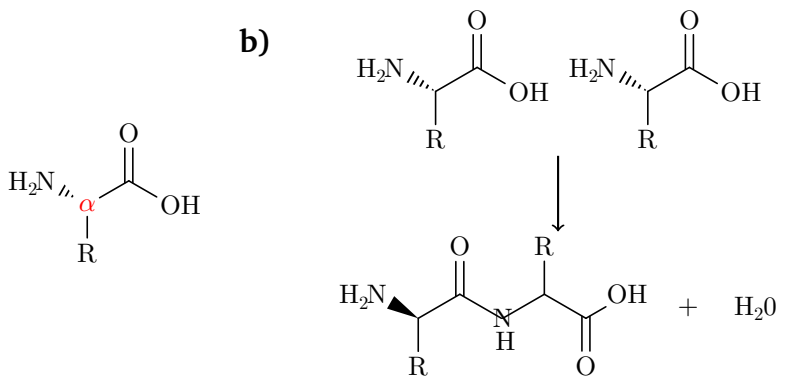

Figure 1.1: Chemical structure of an amino acid. An amino acid consists of an amino group NH_2 and a carboxylgroup CO_2H. The identity of an amino acid is determined by the side chain R.

1.2 Physical and chemical properties of proteins

1.2.1 Amino acids

The building blocks of proteins are amino acids. The central atom of an amino acid is a carbon atom, the α-carbon, to which a carboxyl group (CO_2H), an amino group (NH_2), and a hydrogen atom are bound. The identity of the amino acid is determined by a fourth group, the side chain R, which is bound to the α-carbon (see Fig.1.1.a)). In general, there are twenty different side chains, i.e., twenty different amino acids, which can be chemically modified after translation [3, 4]. The side chains of amino acids can have very different physical and chemical properties. For instance, side chains can differ in size: The side chain of glycine consists of only one hydrogen atom, while the largest amino acid, tryptophan, has a group with a benzol ring.

Since many proteins are found in aqueous environments, an important property of amino acids is their hydrophobicity. Hydrophobic amino acids mostly have non-polar side chains and are predominantly found in the inner part of the protein, where they are not exposed to the surrounding water. Amino acids with a polar or even charged side chain are mostly hydrophilic and are found at the surface of the protein, where they can form energetically favorable bonds with the molecules of the surrounding water. The side chain of proline is special as it is linked back to the amino group, which causes a kink in the backbone of the protein structure.

1.2.2 Protein structure

In the ribosome amino acids are covalently bound to a growing polypeptide chain, the protein. Under the loss of one water molecule the OH group of the carboxyl group of one amino acid is bound to the N_2H group of the next amino acid by the formation of a peptide bond. The amino acid as part of the polypeptide chain is called a residue. The polypeptide chain is very flexible and can attain many different conformations.

Protein structure can be described in a hierarchical scheme. The sequence of amino acids is referred to as the primary structure of the protein. The sequence is written from the N-terminus, where the amino group is not involved in a peptide bond, to the C-terminus, where the carboxyl

name	three letter code	one letter code	interactivity	polar	hydrophobic	small
alanine	ALA	A	0.137		✓	✓
glutamic acid	GLU	E	-0.048	✓		
glutamine	GLN	Q	0.032	✓		
aspartic acid	ASP	D	-0.123	✓		✓
asparagine	ASN	N	-0.035	✓		✓
leucin	LEU	L	0.425		✓	
glycine	GLY	G	-0.046			✓
lysine	LYS	K	-0.010	✓		
serine	SER	S	-0.043	✓		✓
valine	VAL	V	0.408		✓	✓
arginine	ARG	R	0.036	✓	✓	
threonine	THR	T	0.059	✓		✓
proline	PRO	P	0.002		✓	✓
isoleucine	ILE	I	0.417		✓	
metthionine	MET	M	0.175		✓	
phenylalanine	PHE	F	0.408		✓	
tyrosine	TYR	Y	0.317	✓	✓	
cystine	CYS	C	0.275	✓	✓	✓
tryptophan	TRP	W	0.236	✓	✓	
histidine	HIS	H	0.055	✓		

Table 1.1: Standard amino acids with codes and interactivity defined from $U(a, b)$ (table adopted from [4]).

group is free. In a protein sequence, the amino acids are abbreviated by a one letter code (see Table 1.1).

The secondary structure refers to regular periodic folding patterns, which involve residues near to each other in sequence and can be found in almost all proteins. The most important secondary structure elements are α-helices, β-sheets, and β-turns. In α-helices the chain is folded into a right-handed coil (red parts in Fig. 1.2.a)). β-sheets are formed by two or more almost linearly stretched segments of the chain that run parallel or anti-parallel to each other (yellow parts in Fig. 1.2.a)). A β-turn is a sharp bend in the chain that redirects the chain back into the inner part of the protein. A characteristic feature of all secondary structure elements is that they are stabilized by hydrogen bonds. For instance, the α-helix is stabilized by hydrogen bonds between the backbone amide hydrogen and the oxygen of the amino acids that are four residues apart in sequence, but are close to each other in space. The fraction of the chain that is found in a secondary structure varies considerably. The protein keratin, which is the key structural component of hair, consists almost entirely of α-helices, while most parts of the chain of silk, which is also a protein, are folded in β-sheets. Other proteins have no secondary structure at all and the complete chain is folded in a random coil.

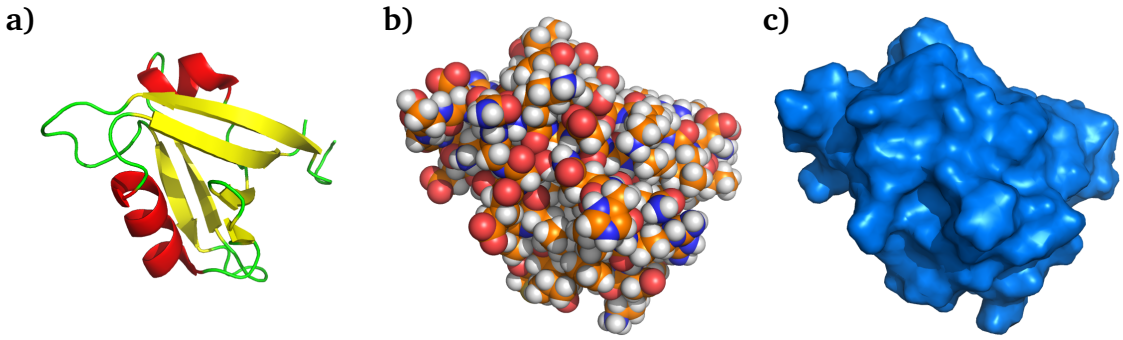

Figure 1.2: Different representations of protein structure (chain with PDB-ID 1m5zA). **a)** Cartoon representation, displaying secondary structure (α-helix (red), β-sheet (yellow), loop(green)), **b)** space filling model (atoms displayed as spheres with van-der-Waals-radius, color code: carbon: organ, hydrogen: gray, nitrogen: blue, oxygen: red), **c)** molecular surface (Connolly surface, which shows the surface accessible to the solvent) (rendered with PyMOL [6]).

The part of the chain that is found in a secondary structure element is very restricted in its possible conformations. Thus, secondary structure elements represent a kind of structural units, which are assembled to a protein structure. Often specific arrangements of secondary structure elements are found in different proteins. Sometimes secondary structural arrangements are repeated yielding highly symmetric protein structures [5].

The tertiary structure defines the overall spatial arrangement of the secondary structure elements as well as other parts of the chain, which are not next to each other in sequence. The three dimensional structure is compact, but not as perfectly compact as a sphere, and shields the inner part, which consists mostly of hydrophobic amino acids, from the solvent (see Fig. 1.2.c)).

Some proteins consist of more than one polypeptide chain, which are bound to each other in a certain arrangement. This arrangement of proteins is referred to as the quarternary structure.

1.3 Protein thermodynamics and design principles

Besides the peptide bond that forms the polypeptide chain, amino acids in a protein structure interact by many different non-covalent interactions. Charged amino acids interact by the laws of electrostatics. As mentioned before, hydrogen bonds between amino acids are found most prominently between amino acids in the same secondary structure element. Salt bridges between positively and negatively charged side chains form a bond, which can be seen as a combination of a hydrogen bond and an ionic bond. All atoms in the protein interact via the Van-der-Waals force and are repelled if they come close to each other in space, giving rise to excluded volume.

An exception is the amino acid cysteine, which forms the only covalent bond, namely to other cysteine residues, which are not next in sequence, by a disulfide bridge. If two connected cysteines are distant in the sequence, the number of misfolded states is dramatically reduced and hence the native state is stabilized.

The driving force of protein folding is the hydrophobic effect [7]. The origin of the hydrophobic effect is the interaction of the amino acids with the surrounding water. Hydrophobic amino

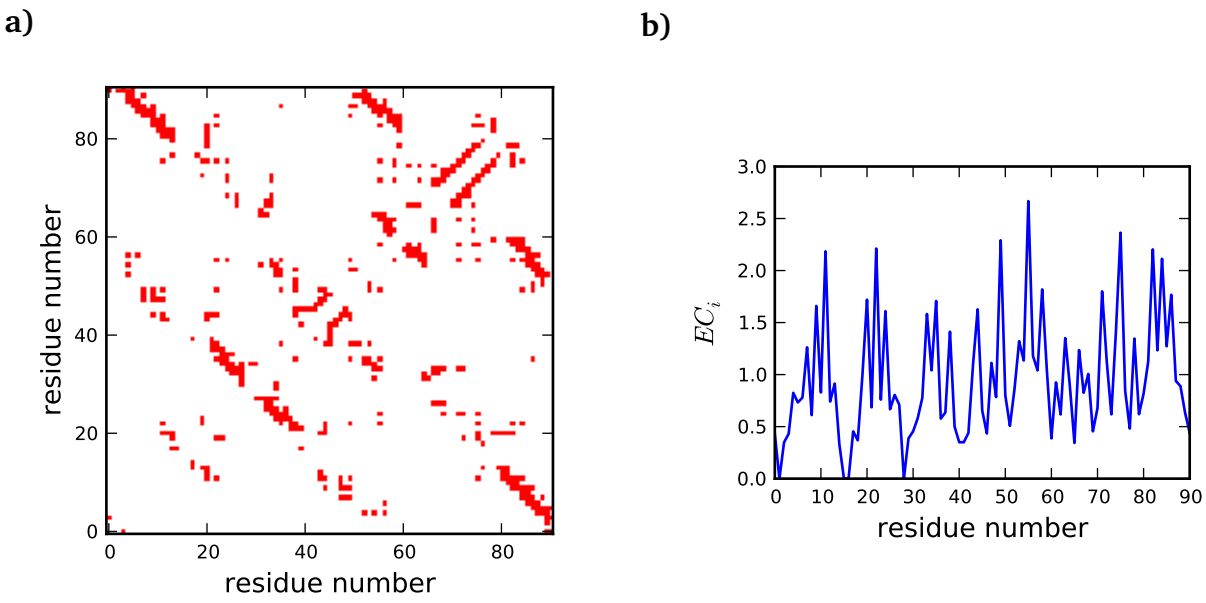

Figure 1.3: Contact matrix (**a**) and Effective Connectivity (**b**) of protein 1m5zA.

acids do not have polar side chains and the surrounding water cannot form favorable hydrogen bonds with the amino acid. Thus, the water builds cages around the amino acid, which have a lower entropy than bulk water. If two hydrophobic amino acids get close to each other they are enclosed in the same cage, which consists of less water molecules than two separate cages. Consequently, the entropy of the water is increased and the free energy is decreased, which results in an effectively attractive interaction [3].

A detailed modeling of all forces – even in a classical treatment – brings about a large computational effort. In particular, an accurate modeling of the interaction with water, which gives rise to important contributions of the proteins folding energetics and dynamics, makes the explicit modeling of the surrounding water necessary. This effort is dramatically increased if the energy of a large ensemble of structures needs to be computed. Therefore, coarse grained models for the folding energetics of proteins were developed. These are not very accurate, but capture the basic principles of protein folding. Apart from the electrostatic interaction between amino acids all interactions are short ranged, which motivates folding models that approximate the folding free energy as a sum of contact interactions between residues.

In these models, the structure is represented by a matrix of dimension $L \times L$, where L is the number of residues in the chain. The entry C_{ij} is one, if the residues i and j are closer than a specific threshold d_{thr} in space,

$$C_{ij} = \begin{cases} 1 & \text{if } d(i,j) < d_{\text{thr}} \\ 0 & \text{else} \end{cases} \quad . \tag{1.1}$$

In this thesis, two residues are defined as being in contact if two non-hydrogen atoms of each of the two amino acids are less than 4.5 Å apart in space. Residues that are less than 3 residues apart in sequence are always in contact and their contacts are therefore neglected. Thus, contacts account not for interaction local in the sequence. Fig. 1.3 shows the contact matrix for the protein chain 1m5zA. Indeed, this representation does not loose relevant information of protein structure as the three dimensional structure of the protein can be reconstructed from the contact

matrix for compactly folded chains [8, 9]. The free energy of a sequence A that is folded into a structure C is given by

$$E(A, C) = \sum_{i<j-2} C_{ij} U(A_i, A_j) \quad . \tag{1.2}$$

The $U(a, b)$ of the interaction matrix depends on the types of amino acids a and b. Such a model is knowledge based, that is, the interaction matrix is inferred from experimentally determined native protein structures and the corresponding protein sequences, which are stored in the Protein Data Bank (PDB) [10]. The values $U(a, b)$ are found from a fit, which maximizes the stability of sequences A^{nat} in their native states C^{nat} with respect to misfolding [11]. Furthermore, the energy model is such that the energy landscape is correlated, that is, the free energy of a structure C should be lower the closer it is to the native state C^{nat}. A correlated energy landscape is seen as an important factor for a fast folding of the protein into its native state. The similarity is measured by the contact overlap $q(C, C^{\mathrm{nat}})$, which measures the fraction of contacts, which the native structure C^{nat} and a misfolded structure C have in common,

$$q(C, C^{\mathrm{nat}}) = \frac{\sum_{i<j} C_{ij}\, C^{\mathrm{nat}}_{ij}}{\max\left(\sum_{i<j} C_{ij}, \sum_{i<j} C^{\mathrm{nat}}_{ij}\right)} \quad . \tag{1.3}$$

This is achieved by finding the interaction energies that maximize the Boltzmann weighted contact overlap (BWCO) with the native structure, where the interaction energy is in units of temperature,

$$\mathrm{BWCO} = \sum_{\{C^{\mathrm{nat}}, A^{\mathrm{nat}}\}} \frac{\sum_{C'=C^{\mathrm{nat}}, \{C\}} e^{E(C', A^{\mathrm{nat}}, U)}\, q(C,\, C^{\mathrm{nat}})}{\sum_{C=C^{\mathrm{nat}}, \{C\}} e^{E(C', A^{\mathrm{nat}}, U)}} \quad , \tag{1.4}$$

where $\sum_{a,b} U^2(a, b)$ is restricted to a constant value, to ensure normalization. The resulting values for $U(a, b)$ are depicted in Fig. 1.4. It can be seen that hydrophobic amino acid have a negative, i.e., attractive, interaction potential, whereas most of the hydrophilic amino acids interact repulsively with other amino acids.

The values $U(a, b)$ have to be interpreted in units of the temperature that is employed in the Boltzmann average in eq. (1.4). Moreover, the interaction parameters $U(a, b)$ are temperature dependent, because the major underlying physical interaction, namely the hydrophobic effect, is temperature dependent. Therefore, the energy model cannot be generalized to other temperatures, which deviate from the optimal growth temperature of the proteins in the fit.

Of course, one might expect contributions to the energy that are due to the local conformation of the chain, which, for instance, is described by the dihedral angles. There exists an extension to the model, which depends on secondary structure [12]. Secondary structure exhibits a particular dihedral angle pattern. Thus, the extension can be seen as an approximated energy function of dihedral angels. However, the contact interaction alone suffices to assign the lowest free energy to the native state. Therefore, local interactions are neglected in this thesis.

A useful approximation of the free energy parameters is given by the hydrophobic approximation, where the interaction matrix is approximated by its eigenvector h to the eigenvalue $\epsilon_H (= -2.625)$ with the largest absolute value,

$$U(a, b) \approx \epsilon_H\, h(a) h(b) \quad , \tag{1.5}$$

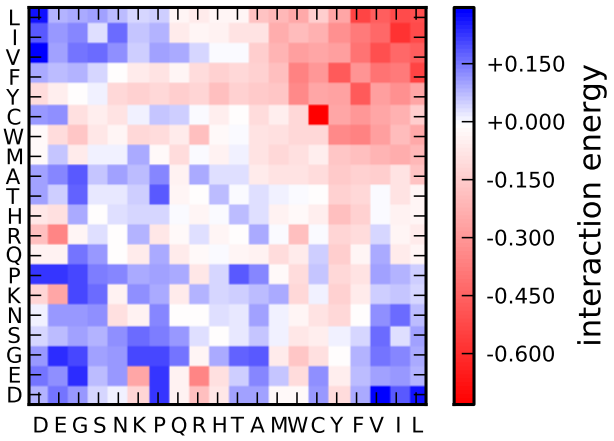

Figure 1.4: Interaction matrix $U(a,b)$ of the contact interactions of residues a and b. Residues are sorted according to their interactivity $h(a)$.

which allows for approximating the free energy in a quadratic form

$$E(A,C) \approx \frac{\epsilon_H}{2} \sum_{i,j} C_{ij}\, h(A_i)\, h(A_j) \quad . \tag{1.6}$$

The vector h was termed the interactivity of an amino acid (values are listed in Table 1.1) and defines a hydrophobicity scale, which is well correlated with other hydrophobicity scales. Thus, the interactivity is used as the hydrophobicity scale in this thesis.

Fig. 1.5 shows that the interaction energy $U(a,b)$ is very well approximated by the eigenvector, i.e., by the product $\epsilon_H\, h(a)h(b)^T$. If the mean interaction energy of a sequence is considered, where positive and negative interaction energies compensate each other, it can be seen that the approximation can be improved by adding a repulsive constant $U_{\text{rep}} \approx 0.04$ (Fig. 1.5.b)).

The quantity which describes the thermodynamic stability of the native state is the difference between the free energy of the native state E_{nat} and the free energy of other conformations of the chain $G_{\text{unfold}} + G_{\text{misfold}}$, which consists of unfolded and misfolded conformations,

$$\Delta G = E_{\text{nat}} - G_{\text{unfold}} - G_{\text{misfold}} \quad . \tag{1.7}$$

The more negative ΔG the more stable is the native state. Interestingly, natural proteins in nature are observed to be only marginally stable, that is, ΔG is only 1-2 kcal/mol below zero, which is in the order of the energy of one hydrogen bond. The reason for the marginal stability is still a topic of discussion. On the one hand, marginal stability is believed to be positively selected as it ensures flexibility of the chain, which might be important for function [13]. In addition, marginally stable proteins are believed to be degraded more easily, which is important for the self-regulation of the cell. On the other hand, it is argued that the high stability of proteins does not need to be negatively selected, and marginal stability is rather an effect of evolution, where mutations that destabilize the native state are more common than mutation that lead to a more stable protein [14, 15].

However, the stability of the native state can be increased where it is needed. For instance, the stability of the native state is more difficult to achieve for organisms that are found in

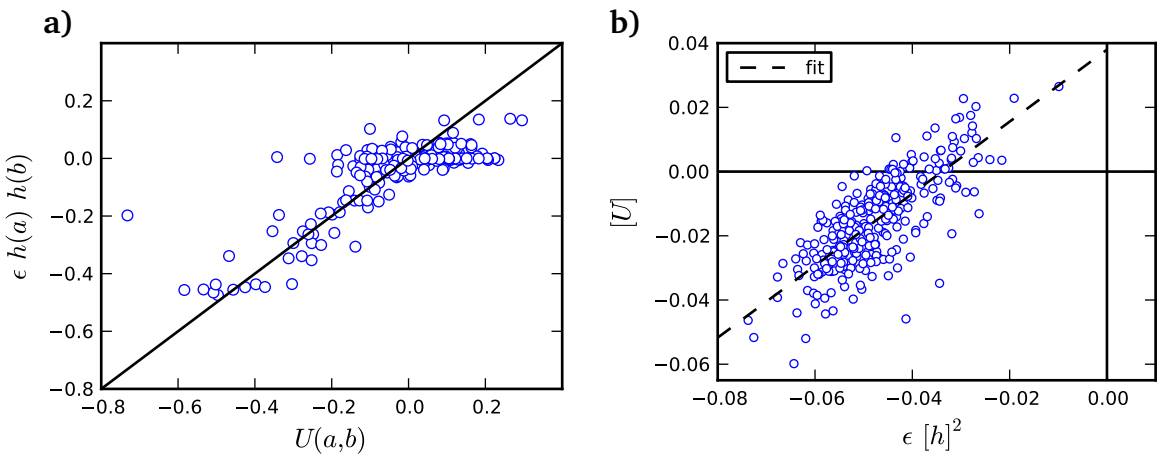

Figure 1.5: Hydrophobic approximation. **a)** The interaction matrix can well be approximated by its principal eigenvector, the interactivity $h(a)$. **b)** The mean interaction $[U]$ of all residues pairs in natural proteins reveals that the hydrophobic approximation can be improved by adding a small repulsive term $U_{\text{rep}} \approx 0.04$.

environments with high temperatures. That is because, according to the Boltzmann distribution, unfolded and misfolded structures get relatively a higher statistical weight with increasing temperature. Nevertheless, the proteins in thermophiles fold into stable structures at high temperatures.

An important question in protein science is how to design a sequence that folds into a certain structure. As the protein's function is determined mainly by its structure, design of a sequence allows for drug design and proteins with tailored functionality. As will be discussed in the next chapter, only E_{nat} and G_{misfold} depend on the amino acid composition on the chain, while G_{unfold} depends only on the length of the chain. Thus, the thermodynamic stability can be improved by either lowering E_{nat} or raising G_{misfold}. The former strategy is termed positive design, while the latter is referred to as negative design, as it reduces the probability for the chain to misfold.

In the terminology of the energy model defined above, positive design is the strategy to make native contacts, i.e., contacts that are formed in the native state, as attractive as possible. The obvious solution is to select hydrophobic amino acids, which have strongly attractive contact interactions. However, such a sequence will also make non-native contacts more attractive, which are formed in misfolded conformations but not in the native state. Consequently, the free energy of the native state E_{nat} will not be significantly lower than the free energy of a misfolded structure. Here comes into play negative design, which ensures that misfolded structures acquire a large energy by making non-native contacts repulsive. To reconcile the two contradicting design principles amino acids from all ranges of hydrophobicity have to be used for sequence design by placing hydrophobic amino acids in the core of the native structure and hydrophilic amino acids on the surface of the protein, where not as many contacts as in the core residues are present.

However, this is only a qualitative view on the design principles. More accurately, the formation of a contact in the misfolded ensemble occurs with different probabilities and is not independent of the formation of other contacts. These statistical properties have an influence on the free energy of the misfolded ensemble. In Chapter 2, a statistical analysis of the misfolded free energy of wild type sequences is used to investigate if wild type sequences show signals of

negative design that stem from the statistical properties of the misfolded ensemble. Moreover, a method for designing sequences is proposed, which explicitly takes into account the statistical properties of the misfolded ensemble.

1.4 Evolution of proteins

The amino acid sequence of proteins is encoded in the sequence of nucleotides in the DNA (deoxyribonucleic acid), which is a sequence built up from the four different nucleotides: guanine, alanine, cytosine, and thymine. The DNA sequence is transcribed into a mRNA sequence, which is identical to the DNA sequence, except that the nucleotide alanine is exchanged to uracil. In the ribosome the mRNA sequence is translated into the protein sequences. Three successive nucleotides, called a codon, encode one amino acid and are matched by the anti-codon of the tRNA, to which the encoded amino acid is attached. The mapping from a codon to an amino acid is unambiguous (the standard genetic code, which is found in most organism is listed in Table A.2 on page 93). The translation stops if the so called stop codon occurs in the mRNA sequence.

Errors in the replication of the DNA or environmental stress will change the nucleotide sequence. These errors are called mutations. In the course of evolution nucleotides are exchanged by mutations or codons are inserted or deleted from the DNA sequence. However, the translated amino acid sequence may stay the same if the codon resulting from a mutation encodes the same amino acid, in which case the mutation is called synonymous, otherwise the mutation is non-synonymous.

Mutations change the phenotype, which is in our case the amino acids sequence of the protein. A change in the protein can have an effect on the fitness of the phenotype, i.e., the ability of the organism to live and procreate. The effect of the mutation on the fitness determines the fade of the mutation. Advantageous mutations, which increase the fitness, have a higher probability to become fixated in the population than deleterious mutations, which reduce the fitness of its carrier.

In the study of protein evolution the important question is which mutations are excepted in a protein coding DNA sequence. The most obvious selection criterion is the ability of the protein to correctly function in the cell. As the structure of proteins is related to its function, the structure imposes constraints on sequence evolution. Thus, while structure is highly conserved during evolution the protein sequence changes more rapidly [16, 17]. In fact, protein sequences, which have diverged during evolution, can have as little as 20% identical amino acids and still fold into very similar structures [18].

The selection pressure on different sites in the proteins can vary significantly and some sites in the protein structure are more conserved than others. Sites that are involved in the function of the protein, e.g., that are located at a binding site, were found to be strongly conserved [19]. The selection pressure from structural constraints varies from site to site. For instance, residues in the core of the protein structures are more conserved than surface residues [20, 21, 22]. In the core amino acids are densely packed and subject to more constraints by the environment of other amino acids. For instance, the smallest amino acid glycine is highly conserved as it can hardly be replaced by larger amino acids without disrupting the structure [4].

The assumption that a protein should always fold into the same native structure allows to formulate a simple model of protein sequence evolution, where the thermodynamic stability of the native state is subject to selection. In this context, the question which sequences fold into

the same structure is of interest. This question is referred to as the *inverse folding* problem. Although the number of sequences is large, it is possible to describe the sequences statistically by the probability $P_i(a)$ to observe an amino acid a at a site i.

The starting point is the hydrophobic approximation, which allows to define an optimal hydrophobicity profile (HP) , which has optimal stability in the native state. The optimal HP h_{opt} is assumed to be the mean hydrophobicity averaged over evolution,

$$h_i^{\text{opt}} = \sum_a h(a)\, P_i(a) \quad . \tag{1.8}$$

For an analytical solvable model it is advantageous to adopt the infinite alphabet approximation, which assumes the hydrophobicity at site i to be a continuous variable h_i. Then, the problem of maximizing the free energy of the native state $E(C^{\text{nat}}, h)$ is equivalent to maximizing the quadratic form,

$$E(C^{\text{nat}}, h) = -\frac{\epsilon_H}{2} \sum_{ij} C_{ij}\, h_i\, h_j \quad . \tag{1.9}$$

To ensure normalization, the optimal HP has to be subjected to constraints. If the mean square $\left[h^2\right]$ of the optimal HP is restricted, the solution is the eigenvector to the largest eigenvalue (principal eigenvector (PE)) of the contact matrix C^{nat}, which was studied before [23].

A profile that is related to the PE is the Effective Connectivity (EC), which results from maximization of the quadratic form under the constraints $\left[h\right] = 1$ and $\left[h^2\right] = B > 1$ [24]. The parameter B is set to $\left[c^2\right]/\left[c\right]^2$, where the vector entry c_i is the number of contacts of residue i. The constraints on the EC are analogous to restricting the hydrophobicity and therefore the misfolded free energy (for a more detailed discussion see Section 2.4.5). It was found that the EC correlates better with the HP of wild type sequences than the PE and has not the problems of the PE that arise for multi-domain proteins, where the PE describes only one domain, and is effectively zero for all other domains.

The EC can be written as

$$x_i = \frac{1}{W} \sum_j \frac{1}{\nu^{(j)} - \Lambda}\, v_i^{(j)} \quad , \tag{1.10}$$

where $v^{(j)}$ is the jth eigenvector of the contact matrix that is rescaled such that its mean value $\left[v\right]$ is one. W and Λ are Lagrange parameters, which are to be adjusted to meet the constraints. For small globular structures Λ is the nearest to the largest eigenvalue $\max_j \nu_j$ and the EC is therefore very similar to the eigenvector to the largest eigenvalue. The EC of the protein 1m5zA is shown in Fig. 1.3.b).

The distribution of amino acids at a site i can be found by applying the principle of maximal entropy, which gives rise to a Boltzmann distribution,

$$P_i(a) = \frac{n_c(a)}{Z(\beta_i)} \exp\left(-\beta_i\, h(a)\right) \quad , \tag{1.11}$$

where $n_c(a)$ is the number of codons that code amino acids a, which is proportional to the background frequency of the amino acid a due to the mutation process alone. $Z(\beta_i) = \sum_a \exp(-\beta_i h(a))$ is the partition function, which ensures normalization. The temperature β_i

has to be determined from the mean h_i^{opt}. Since the EC is assumed to be optimally correlated with h^{opt}, the temperature β_i has to be a function of the value of the EC at site i. From simulation of protein sequence evolution and a statistical analysis of proteins, the Boltzmann distribution and the functional relationship between the temperature and the EC were confirmed [24].

As the EC is computed from the protein structure, it also represents structure. Thus, the EC connects protein structure and sequence evolution. This relationship was exploited to reveal evolutionary related proteins from structural and evolutionary information within one alignment framework [25, 26]. Moreover, the EC was successfully applied in protein structure prediction [27]. In Section 2.4.5, a variant of the EC is proposed that respects more details of the misfolded ensemble.

However, different sites in a protein do not evolve independently. The decrease in stability by a mutation at one site can be compensated by a mutation at another site. These compensatory mutations are abundant in natural proteins [13, 28, 29]. Indeed, the ability of a protein to fold into a unique structure is the result of the interaction of all residues. Consequently, strong correlations between the substitutions of amino acids at different sites are expected. Furthermore, the observation of correlated mutations was useful for the prediction of protein structure and protein interacting interfaces by predicting pairs of residues that are close in space and for revealing path ways of signal transduction through the protein structure by strongly coupled amino acids. However, no model exists that describes correlated mutations quantitatively. In Chapter 3, the maximum entropy model that was discussed for independent sites $P_i(a)$ is extended to pair specific amino acid probabilities $P_{ij}(a, b)$.

2 Negative Design

2.1 Introduction

As laid out in the introductory chapter, the correct functioning of a protein requires the native protein structure to be thermodynamically stable, that is, the chain has to be stable against unfolding and misfolding. While most of the research concentrated on the decrease of E_{nat} by positive design, negative design has attracted increased attention only recently. Particular attention attracted the problem by the observation of correlated mutations of sites that are distant in the native state. Besides functional explanations, which will be discussed in Chapter 3, one explanation is that the interaction of these residues in a non-native state influence the stability of the native state and thus are subject to selection.

From studying lattice proteins, which represent proteins as a walk on a cubic lattice, Horovitz *et al.* found that the effect of a pair of residues on stability is stronger the larger the contact frequency in the Boltzmann ensemble of this pair is. They concluded that the stability of a fold can be increased in two ways: First, a native contact that is rarely formed in the misfolded ensemble can be made more attractive, which has little effect on the misfolded free energy. Second, a non-native contact, which is frequently formed in the misfolded ensemble can increase the stability of the fold if it is made repulsive [30].

In thermophiles, the thermodynamic stability of protein structures is more demanding to achieve. Based on their finding that negative design is enhanced in proteins with a high average contact frequency of non-native contacts, Horovitz *et al.* predicted that thermophiles are particularly optimized with respect to misfolding. In accordance with this prediction, Berezovsky *et al.* found that sequences of thermophiles show an amino acid composition that enhances positive as well as negative design: They observed that the frequency of extremely hydrophobic as well as of charged, i.e., hydrophilic, amino acids in the sequences increases with the optimal growth temperature of the organism [31].

These results show that negative design has important implications for protein folding and evolution. However, the theoretical modeling of the misfolded free energy is complicated by the enormous number of misfolded conformations. Therefore, simple approximating schemes have been developed that assume contacts as independent random variables. This is, however, only an inaccurate description of the misfolded ensemble. The contact frequency in the Boltzmann ensemble, which has been identified as a key variable for negative design, depends on the frequency and correlation of contacts, which are formed in a misfolded structure.

In the following, a description of the free energy of the misfolded ensemble is introduced, which explicitly takes into account contact frequency and contact correlation. This formalism is used to search for evidence that wild type sequences are optimized with respect to negative design. To this end, correlations of sequence features with contact frequency and correlations are investigated.

Design principles are also important in the rational design of sequences. The detailed description of the misfolded free energy allows for designing sequences that fold into a particular structures. To this end, variants of the EC will be discussed, which represent an optimal hydrophobicity profile, which is designed to respect the needs of negative design.

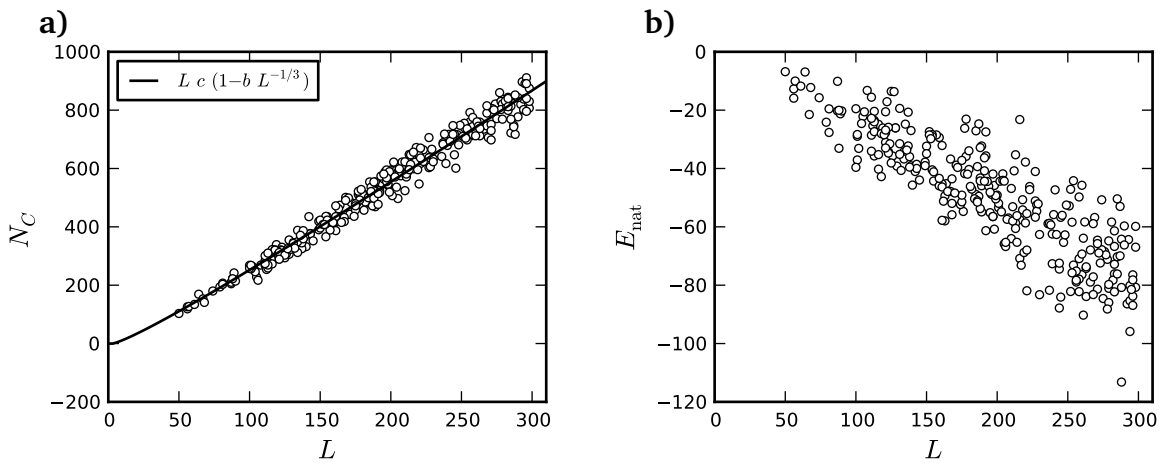

Figure 2.1: Number of **a)** native contacts and **b)** native energy versus length

2.2 Native structure and ensemble of alternative structures

2.2.1 Native structure

Native structures of proteins are compact, however not as compact as a sphere, and have, compared to other compact structures, a very low free energy. In our model the free energy is computed by summing over contacts. The number of contacts in a native structure depends strongly on length and can be fitted very well by $R\,c\,(1-b\,R^{-1/3})$, finding ($c = 3.728$, $b = 1.501$) (see Fig. 2.1.a)). The scaling reflects that every amino acid has on average c neighbors, while residues which are at the surface of the structure, have fewer contacts. The fraction of surface residues is proportional to $N_C^{1/3}$. The value for the fit parameter b nicely agrees with the expectation $b = 1.5$ for a compact self-avoiding walk on a lattice [32].

The free energy is given by

$$E_{\rm nat} = \sum_{i<j-2} C_{ij}^{\rm nat} U(A_i, A_j) \tag{2.1}$$

and decreases with the chain length (see Fig. 2.1.b)). The variation of the free energy is much larger than the variation of the number of contacts, due to the large variety of contact interactions.

2.2.2 Unfolded and misfolded ensemble

Unfolded ensemble

The unfolded state has only very few contacts, if any. Thus, contacts are assumed not to contribute to the free energy, which then is only determined by the conformational entropy, i.e., the logarithm of the number of unfolded states. Under the assumption that the backbone of the chain can be described as a (self avoiding) random walk through space and the side chains of each residue has the same amount of conformations, one finds an exponential scaling of the

number of unfolded states with the length of the chain. That is, the conformational entropy is proportional to the chain length. In consequence, the free energy of the unfolded ensemble is estimated as $G_{\text{unfold}} = s_U\, L\, T$.

Before, the scaling factor s_U was estimated from the free energy difference between the unfolded and native state of proteins, which are known as two state folders, i.e., only the unfolded and the native state are prominent in the ensemble, whereas misfolded states are not. Assuming that all free energy differences were determined at the same temperature, a two parameter fit can be performed to the experimental data,

$$\Delta G/L = a\, E_{\text{nat}}/L - s_U \quad , \tag{2.2}$$

where a is a factor that mediates between the scale of the interaction parameters and units of energy from empirical measurements and s_U is the scaling factor of the conformational entropy of the unfolded ensemble. From the fit to 44 proteins the values $a = 0.485\,\text{kcal/mol}$ and $s_U = 0.0636\,\text{cal/mol}$ were found (private communication, Ugo Bastolla).

The values for the conformational entropy, however, is much less than the expected entropy that is solely due to side chains, i.e., neglecting degrees of freedom of the backbone [32]. Since there is no better estimate, in this work the room temperature is set to 1.2 units of energy and s_U is set to 0.13.

Misfolded ensemble

Misfolded structures are as compact as native structures, i.e., the number of contacts scales with length similarly to native structures. In principle, the misfolded structures are represented by a set of contact matrices $\{C\}$ of compact structures that comply with steric constraints of a compactly folded chain.

Compact structures also favor the formation of secondary structure [7, 33, 34, 35], in agreement with the observation that compactness is more easily attained if parts of the chain exhibit a regular folding pattern. A description of misfolded conformations, which respect steric conditions, compactness, and secondary structure, is difficult. However, many native protein structures with a large variety of different fold topologies and secondary structure content are at hand, which can serve as templates for misfolded structures, provided that they are sufficiently different from the native structure.

To this end, I use a non-redundant set of approximately 1,000 native protein structures from different protein families and of different secondary structure content, with chain lengths ranging from 30 to 1,000 residues. The ensemble of misfolded structures is generated by a procedure known as *gapless threading*, which is usually employed in structure prediction [36]. A chain of a candidate misfolded structure is in general not as long as the length L of the chain, for which the misfolded structure is desired (hereafter referred to as the query). If it is shorter, the misfolded structure is rejected. If the chain is longer, the query is threaded along the backbone of the candidate structure. In other words, all substructures of the candidate structure are considered, which start at residue i and end at residue $i + L - 1$. A candidate structure, whose chain is L_2 residues long, can thus produce $L_2 - L + 1$ substructures, starting from $i = 0$ to $i = L_2 - L$. Threading produces almost 200,000 substructures for a query length of 50 residues. This value quickly drops with increasing query length, until it reaches 40,000 structures for a query length of 300 residues.

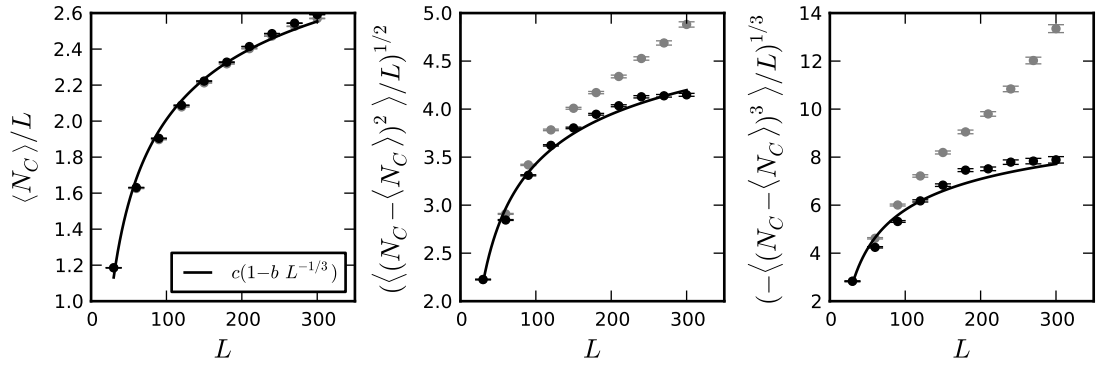

Figure 2.2: Length scaling of moments of contact number in misfolded structures (black: non-compact structures removed, gray: all structures).

order	c_n	b_n
1	3.786	2.177
2	5.935	1.957
3	12.041	2.404

Table 2.1: Fit parameters for length scaling of contact number moments, according to equation (2.3).

Some structures of long chains can comprise two domains of the candidate structure. In this case, the substructure is most likely not compact. Indeed, one can observe outliers which have significantly fewer contacts than the mean number of contacts expected for the respective length (see Fig. A.1). In order to keep the properties of misfolded structures well defined, non compact structures are removed from the set, if their number of contacts is more than three standard deviations below the mean number of contacts expected for the length of the query. This filtering has not a large impact on the free energy of the misfolded ensemble, because the energy of non-compact structures is higher than the free energy of more compact structures.

With the non-compact structures removed, the length dependence of the moments of the contact number can be described by simple and general scaling behavior,

$$\left\langle N_C \right\rangle_L \approx L(c_1(1 - b_1\, L^{-1/3})) \tag{2.3a}$$

$$\left\langle (N_C - \left\langle N_C \right\rangle_L)^2 \right\rangle_L \approx L(c_2(1 - b_2\, L^{-1/3}))^2 \tag{2.3b}$$

$$\left\langle (N_C - \left\langle N_C \right\rangle_L)^3 \right\rangle_L \approx L(c_3(1 - b_3\, L^{-1/3}))^3 \quad , \tag{2.3c}$$

where the fit parameters c_i and b_i are listed in Table 2.1. The filtering of non-compact structures reduces the standard deviation, making the third centered moment less negative (compare black to gray data points in Fig. 2.2).

Since all candidate structures are native protein structures, they can be very similar to the native structure of a wild type sequence, if both, the candidate and the query, belong to the same protein family. This behavior of threading is desired in structure prediction, but is a nuisance here. The similarity of a substructure C to the native state C^{nat} is measured by the

contact overlap $q(C, C^{\text{nat}})$, which is defined as the fraction of contacts that two structures have in common in relation to the maximal number of contacts of both structures,

$$q(C, C^{\text{nat}}) = \frac{\sum_{i<j-2} C_{ij}\, C_{ij}^{\text{nat}}}{\max(\sum_{i<j-2} C_{ij}, \sum_{i<j-2} C_{ij}^{\text{nat}})} \quad . \tag{2.4}$$

Substructures with a contact overlap of greater than 0.5 are considered as similar to the native state and therefore rejected. This filtering is particularly important if the native state has many helical contacts, as two structures with many helices often have a high contact overlap, which gives rise to very negative misfolded free energies.

Using standard statistical mechanics, the free energy of the misfolded ensemble is computed from the partition function of the misfolded ensemble Z_{misfold},

$$G_{\text{misfold}} = -k_B\, T \ln Z_{\text{misfold}} = -k_B\, T \ln \left(\sum_C \exp\left(-\frac{E(C,A)}{k_B\, T} \right) \right) \quad . \tag{2.5}$$

The partition function Z_{misfold} can be computed by summing over all substructures in our threading set. However, threading produces only a subset of all misfolded structures. Hence, the conformational entropy of the misfolded ensemble, i.e., the number of misfolded structures, is underestimated by threading. In particular, the number of misfolded structures N_{thr} produced by threading decreases with increasing chain length of the query, while, similar to the unfolded ensemble, the number of misfolded structures is expected to grow exponentially with length.

This wrong scaling of the conformational entropy can be corrected by subtracting the entropic term of the threading ensemble $T \ln N_{\text{thr}}$ from the free energy and adding an entropy $T\, s_C\, R$, which is proportional to chain length. This correction, however, mends the problem of threading only at higher temperatures. At low temperature only the structure of lowest energy is relevant, which is always overestimated by threading.

The conformational entropy of compact structures was estimated before from the number of arrangements of secondary structure elements, finding $s_{\text{misfold}} = 0.1$.

To assess the stability of the native state, which is the sum of free energy of the misfolded and unfolded ensemble, the free energy of the denatured state has to be considered. Adding the free energy of the unfolded ensemble to the misfolded free energy is analogous to adding the entropic terms $s_C = s_{\text{unfold}} + s_{\text{misfold}} = 0.23$, as the conformational entropy of both ensembles exhibit the same length scaling. However, the combined estimate of the conformational entropy yields unstable long proteins. Therefore, the value of $s_C = 0.1$, which corresponds to the conformational entropy of the unfolded and misfolded ensemble, is adopted for the computations in this work.

2.3 Free energy expansion and Random Energy Model

2.3.1 Theory

While threading represents the gold standard for computing the energy of the misfolded ensemble, it has two major draw backs: It is computationally expensive and provides no analytical

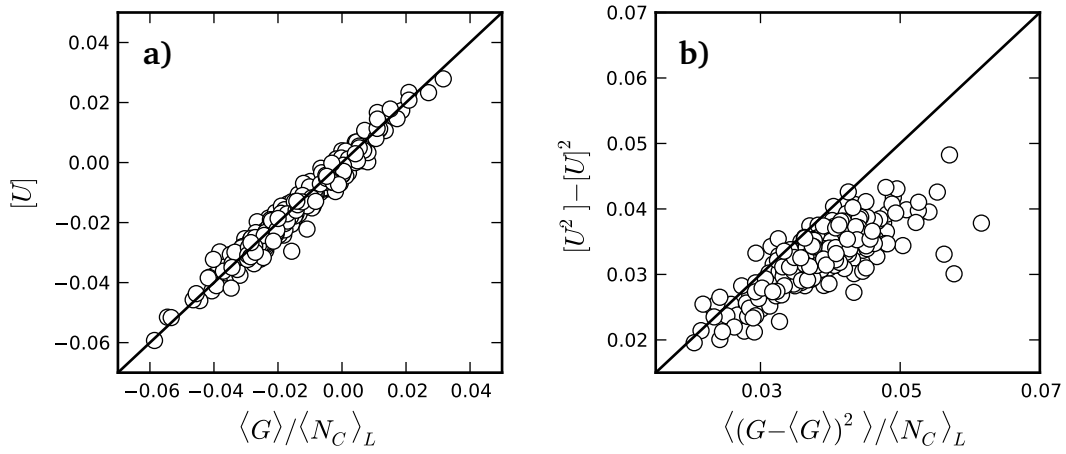

Figure 2.3: REM prediction of the first **a)** and second cumulant **b)**

insight into the properties of the misfolded ensemble. A deeper insight is gained by a cumulant expansion of the free energy of misfolded ensemble (for a derivation see [37]),

$$\begin{aligned} G_{\text{misfold}} \equiv & -k_{\text{B}}T \log\left(\sum_C \mathrm{e}^{-E(C)/k_{\text{B}}T}\right) \\ \approx & \langle E\rangle - \frac{1}{2\,k_{\text{B}}T}\left\langle (E-\langle E\rangle)^2\right\rangle + \frac{1}{6\,(k_{\text{B}}T)^2}\left\langle (E-\langle E\rangle)^3\right\rangle \\ & - \frac{1}{24\,(k_B T)^3}\left[\left\langle (E-\langle E\rangle)^4\right\rangle - 3\left\langle (E-\langle E\rangle)^2\right\rangle^2\right] - k_{\text{B}}T\,S_C \quad . \end{aligned} \tag{2.6}$$

The cumulant expansion is shown here up to the fourth order. The first three cumulants are the mean misfolded energy $\langle E\rangle$ and the second and third centered moment $\langle (E-\langle E\rangle)^n\rangle$. The cumulants can be computed from the distribution of misfolded energies generated by threading. However, methods like threading, that sample chain conformations, are computational expensive. Therefore, simple approximations are needed. The most prominent approximation is the Random Energy Model (REM), which was first introduced for spin glasses [38] and was adapted for the protein folding problem [39, 40]. The REM assumes that the free energies of different structures are independent random variables and follow a Gaussian probability distribution. That is, only the first two cumulants contribute to the cumulant expansion, which then becomes exact. Often it is assumed that contacts are independent random variables, which occur with equal probability. Under the assumption that a contact occurs with a probability $\langle N_C\rangle_L / N_P$, the first two cumulants are easily computed as,

$$\langle E\rangle_{\text{REM}} = \langle N_C\rangle_L \left[U\right] \tag{2.7a}$$

$$\left\langle (E-\langle E\rangle)^2\right\rangle_{\text{REM}} = \langle N_C\rangle_L \left(\left[U^2\right] - \left[U\right]^2\right) \tag{2.7b}$$

Here, squared parenthesis are used to denote the average over all pairs in a sequence, i.e. $\left[U\right] = \sum_{i<j-2} U_{ij} / \sum_{i<j-2}$.

The first two cumulants estimated from the REM are compared in Fig. 2.3 to the cumulants computed by threading. While the first cumulant is very well approximated, the REM estimate

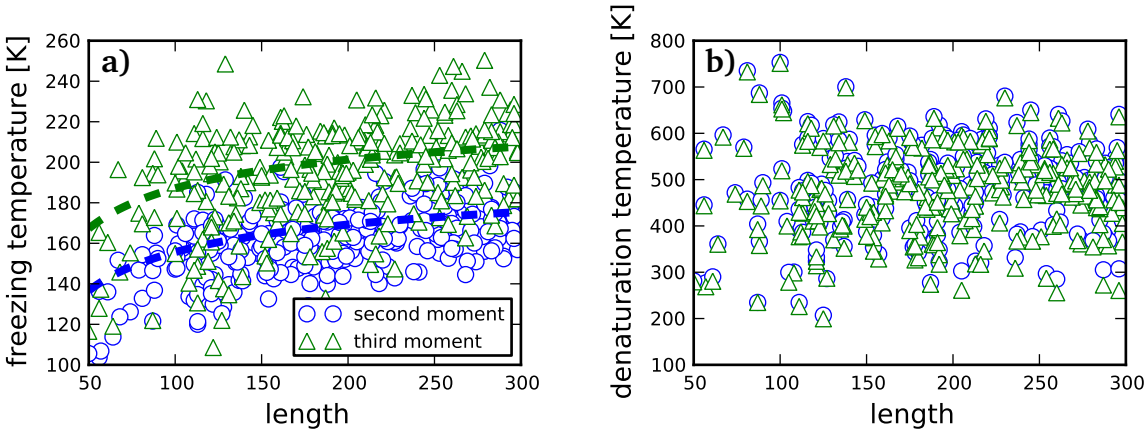

Figure 2.4: a) Freezing temperature versus length computed from the second and third moment of the cumulant expansion. The dotted lines show the freezing temperature computed from the fitted mean cumulants of random sequences. The negative third cumulant increases the freezing temperature. **b)** Denaturation temperature versus length. The denaturation temperature T_{denat} is computed from the condition $\Delta G = E_{\text{nat}} - G_{\text{misfold}}(T_{\text{denat}}) = 0$.

of the second cumulant clearly underestimates the second cumulant generated from threading and is less well correlated than the first cumulant. This deviation increases for higher cumulants, which are not very well approximated by the REM (data not shown).

While the cumulant expansion truncated at a certain order is a good approximation at high temperatures, it exhibits a non-physical behavior at low temperatures as the expansion diverges when the temperature approaches zero. This behavior conflicts with two properties of the free energy. The free energy, as a function of temperature, is monotonously decreasing concave. Since the entropy S is defined as the derivative of the free energy

$$S_{\text{misfold}} = -\frac{\partial G_{\text{misfold}}}{\partial T} \tag{2.8}$$

these conditions are equivalent to imposing that entropy of the misfolded ensemble is larger than zero and increases with temperature.

2.3.2 Freezing transition and folding temperature

The expansion up to the second cumulant always has an extremum, at which the entropy is zero. This point was identified as a glass transition, below which the misfolded ensemble is dominated by only a few structures [41]. Even though the kinetic properties of protein folding are not accessible in equilibrium thermodynamics, as it is done here, it is reasonable to assume that below glass temperature the protein has a largely increased probability to get trapped in a local minimum, i.e., a misfolded structure, during folding. Thus, the freezing temperature can serve as an indicator for the misfolding propensity of a chain.

An analytic expression for the freezing temperature T_C can be found by setting the first derivative of the cumulant expansion up to the second cumulant to zero, which yields

$$T_C = \sqrt{\frac{\left\langle (E - \langle E \rangle)^2 \right\rangle}{S_C}} \approx \sqrt{\frac{\epsilon_2 \, c_1 (1 - b_1 \, L^{-1/3})}{s_C}} \quad . \tag{2.9}$$

Plot 2.4 depicts freezing temperatures computed from expansion up to the second and third order, where the cumulants are computed by threading. The lines are prediction of random sequences from the mean cumulants of random sequences, which has the length scaling $\epsilon_i \, \langle N_C \rangle$ (cf. Section 2.4.3). Since the length scaling is determined by the ratio of the second cumulant, $\epsilon_2 \, \langle N_C \rangle_L$, and the conformational entropy, $s_C \, L$, the freezing temperature increases with length (right hand side of 2.9). Therefore, longer chains are more prone to get trapped in a misfolded structure.

If the third cumulant is considered, the freezing temperature is increased, as the third cumulant is negative for most sequences, which means that the distribution is skewed towards negative values and therefore has a longer tail with negative energies.

The free energy expanded to the third moment might not have a maximum if the third cumulant is positive. This is the case only for a few sequences, for which the second derivative with respect to temperature has a null, which then serves as the freezing temperature. This condition ensures that the entropy, i.e., the first derivative of the free energy, increases with temperature.

For completeness, the denaturation temperature is shown in Fig. 2.4.b), at which the native state becomes unstable, i.e., the free energy of the native state E_{nat} is equal to the free energy of the misfolded state G_{misfold}. Note that the model denaturation temperatures are much higher than typical values of measured denaturation temperatures. That is not only due an uncertainty in the room temperature, but also the temperature dependence of the interaction matrix $U(a, b)$ was neglected.

For good folding properties, which means that the protein does not get trapped in a misfolded state and the native state is stable, the biological temperature T has to be between the freezing and denaturation temperature, $T_{\text{denat}} < T < T_{\text{denat}}$. Under the assumption that the biological temperature is room temperature, this condition is fulfilled for almost all chains.

2.3.3 Testing the cumulant expansion

Here, I assess the cumulant expansion at temperature $T = 1.2$, which is approximately room temperature, and at the freezing temperature (see Fig. 2.5). The free energy is computed by threading with the entropy correction and from the cumulant expansion, using the cumulants that are estimated by threading. Since room temperature is significantly above the freezing temperature, the free energy is well approximated by the first cumulant. The approximation improves if the second and third cumulant are taken into account. If the fourth cumulant is added the error of the approximation increases slightly on average (data not shown). Thus, the expectation is that the fourth cumulant does not contribute significantly to the free energy.

Interestingly, the free energy evaluated at the freezing temperature is identical to the estimate of an extreme value of a Gaussian distribution, which was used before to estimate E_{min} [32],

$$G_{\text{misfold}}(T = T_{\text{freez}}) = \langle E \rangle - \sqrt{\left\langle (E - \langle E \rangle)^2 \right\rangle^{1/2} \, 2 s_C \, L} \approx E_{\text{min}} \quad . \tag{2.10}$$

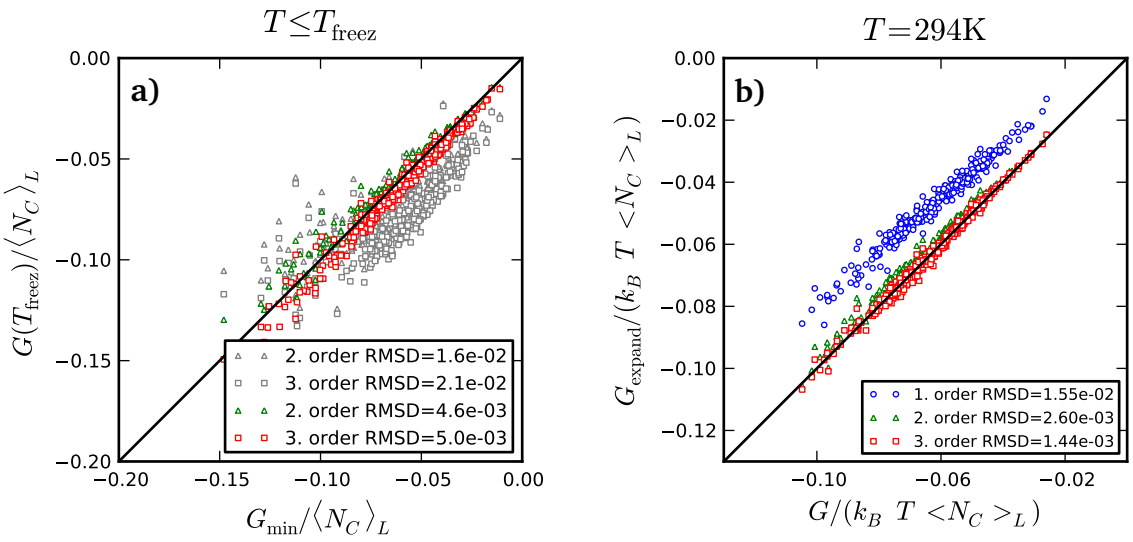

Figure 2.5: Test of the cumulant expansion of the free energy. **a)** Free energy expansion below freezing temperature compared to $E_{\min}$. The conformational entropy is set to $\ln N_{\text{thr}}$ (colored) and the entropy $s_C R$ (gray). **b)** The free energy per contact in units of temperature computed from threading at $T = 294K$ versus the free energy computed from the cumulant expansion, using first cumulant (blue points), first and second (green points), and first to third cumulant (red points). The error decreases with increasing order of the expansion.

Between freezing temperature and zero temperature the free energy of the cumulant expansion is constant. Thus, I compare the free energy of the expansion below the freezing temperature to free energy obtained by threading at zero temperature, which is identical to the free energy of the threading substructure with minimal free energy. For a more meaningful comparison I have to take the logarithm of the number of threading structures as conformational entropy, since the minimal value of the free energy is always overestimated by threading. The approximation of the minimal energy by free energy expansion at the freezing temperature is good for the second order and does not improve with the third order. If the model conformational entropy is taken, which is larger than the number of threading substructures for sequences longer than 120 residues, the minimal energy is always underestimated by the expansion (gray points in Fig. 2.5.a)). To assess the minimal energy, the model conformational entropy will be adopted from here on and the free energy will be computed by the expansion rather than by threading.

2.3.4 Effect of sequence composition

In the REM the misfolded energy depends solely on the sequence composition in terms of the moments of the mean interaction energy $\left[U^n\right]$. Negative design is then defined over the sequence composition, i.e., the amount of hydrophobic and hydrophilic amino acids in the sequence. The free energy of the native state is lowered if more hydrophobic sequences occur in the sequence. However, then the free energy of misfolded conformations is also decreased, which can be compensated by increasing the content of hydrophilic amino acids. Therefore, one

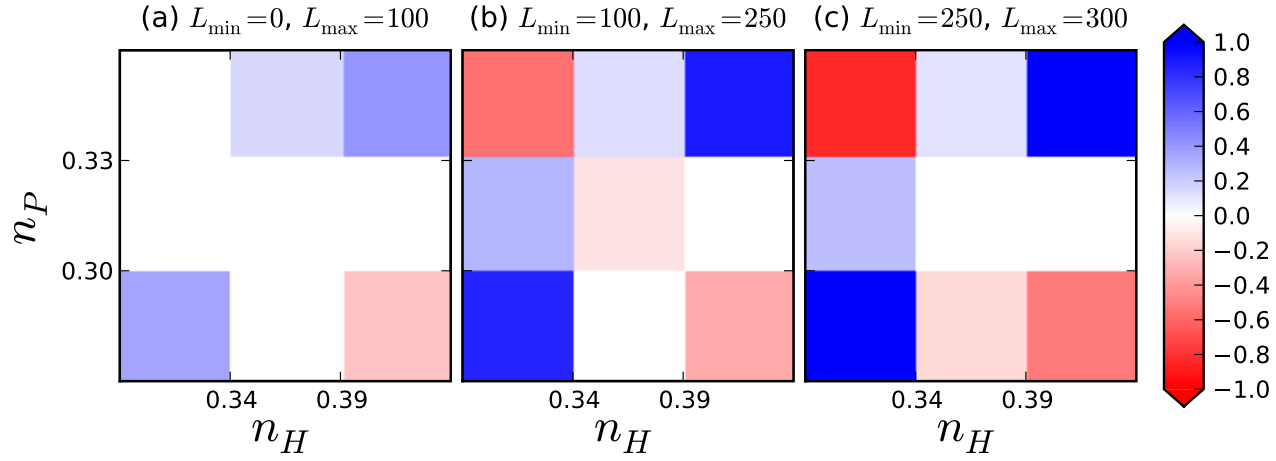

Figure 2.6: Propensity difference of hydrophobic and hydrophilic amino acid content between wild type and random sequences for different sequence lengths. Only propensity differences with a Z-score greater than 3 are shown.

might expect a positive correlation of the content of amino acids for wild type sequences from both ends of the hydrophobicity scale.

With increasing temperature the native state is destabilized, however, it can be stabilized by increasing the content of hydrophobic amino acids. Indeed, it was observed that the content of hydrophobic amino acids increases with the optimal growth temperature of the organism. At the same time the content of charged amino acids, which help to increase the energy of misfolded conformations, increases with temperature [31].

Over a set of 11,000 non-redundant protein sequences I compute the probability of finding the six most hydrophobic (C, Y, F, V, I, L) and most hydrophilic (D, E, G, S, N, K) amino acids in a sequence. The correlation of these two types of amino acids is assessed by the log-propensity $\mathrm{prop}(n_H, n_P)$ of finding a fraction of n_H hydrophobic amino acids and n_P hydrophilic amino acids,

$$\mathrm{prop}(n_H, n_P) = \ln\left(\frac{P(n_H, n_P)}{P(n_H)P(n_P)}\right) \quad . \tag{2.11}$$

The fraction of the two amino acid types is partitioned into three bins, such that the probability to observe a fraction is equal for each bin. For random sequences the propensity has a non-zero expectation value as n_H and n_P are strongly anti-correlated, because of the normalization of the amino acid fractions. Therefore, the average log-propensity for random sequences is subtracted from the propensity for wild type sequences. The resulting difference is displayed in Fig. 2.6. In order to assess the significance of the result, the standard deviation of the log-propensity from 100 samples of random sequences is measured and deviations from random sequences, which are less than three standard deviations apart from the propensity of random sequences, are color coded as white. Clearly, a high content of hydrophobic amino acids is correlated with a high content of hydrophilic residues. As expected, the deviation from random sequences increases for longer proteins.

2.4 Beyond Random Energy Model: Structural correlations

2.4.1 Theory

As discussed above, cumulants are only approximately described by the REM. Indeed, it was argued before that the correlations in the conformational ensemble are critical to protein folding [42]. Those correlations can be expressed as correlations between contacts, which have a significant contribution to the cumulants,

$$\langle E \rangle = \sum_{i<j} \left\langle C_{ij} \right\rangle U_{ij} \tag{2.12a}$$

$$\left\langle (E - \langle E \rangle)^2 \right\rangle = \left\langle \left(\sum_{i<j} \left(C_{ij} - \left\langle C_{ij} \right\rangle \right) U_{ij} \right)^2 \right\rangle = \sum_{i<j} \sum_{k<l} \left(\left\langle C_{ij} C_{kl} \right\rangle - \left\langle C_{ij} \right\rangle \langle C_{kl} \rangle \right) U_{ij} U_{kl} \tag{2.12b}$$

$$\left\langle (E - \langle E \rangle)^3 \right\rangle = \sum_{i<j} \sum_{k<l} \sum_{m<n} \left\langle \left(C_{ij} - \left\langle C_{ij} \right\rangle \right) \left(C_{kl} - \langle C_{kl} \rangle \right) \left(C_{mn} - \langle C_{mn} \rangle \right) \right\rangle U_{ij} U_{kl} U_{mn} \tag{2.12c}$$

where the brackets $\langle . \rangle$ denote the average over misfolded conformations. The first cumulant is a sum over all pairs of residues, the contribution of the pair (i, j) is weighted with its average contact frequency $\left\langle C_{ij} \right\rangle$ that is expected in a misfolded conformation. The second cumulant depends on the covariance tensor $\left\langle C_{ij} C_{kl} \right\rangle - \left\langle C_{ij} \right\rangle \langle C_{kl} \rangle$ of the contacts (i, j) and (k, l), and will be abbreviated by S_{ijkl}. Within the REM contacts are assumed not to be correlated, i.e., the contact correlation tensor could be written as $\delta_{i,k} \delta_{j,l} \langle N_C \rangle_L / N_P$. Consequently, in the REM the sequences with the same sequence composition produce the same misfolded free energy. With contact correlation not only the sequence composition but also the positional correlation of amino acids in the sequence becomes important.

Contact correlations are connected to the moments of the contact number by normalization. For instance, for the contact frequency and contact correlation one finds the normalization conditions

$$\sum_{i<j}^{L} \left\langle C_{ij} \right\rangle = \langle N_C \rangle_L \tag{2.13a}$$

$$\sum_{i<j}^{L} \sum_{k<l}^{L} S_{ijkl} = \left\langle \left(N_C - \langle N_C \rangle_L \right)^2 \right\rangle_L \quad . \tag{2.13b}$$

Measuring contact frequency and correlations

For a better understanding of the properties of the misfolded ensemble, it is interesting to characterize contact frequency and correlations. As the cumulants are computed by threading, the average contact frequency and the contact correlation can be measured using threading as well. To measure the contact frequency, I average the contact matrices of the threading substructures using different window sizes for threading. In Fig. 2.7 the contact frequency $\left\langle C_{ij} \right\rangle$ is plotted

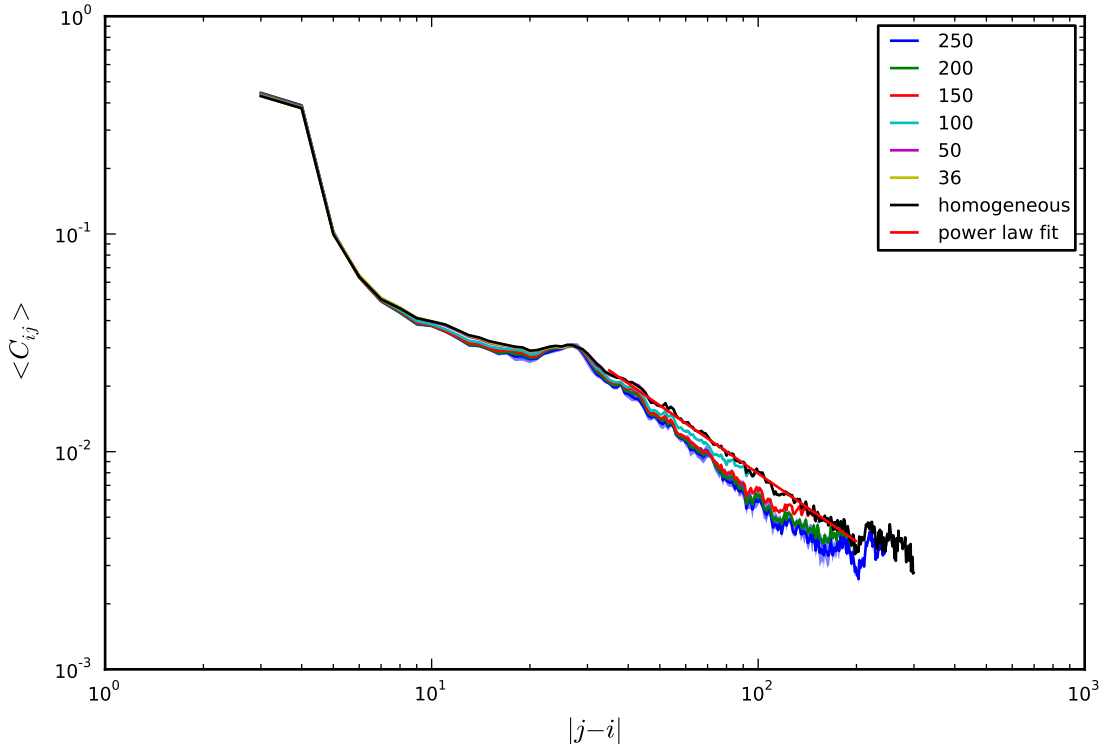

Figure 2.7: Contact frequency measured with threading and within homogeneous approximation versus distance in sequence. The contact frequency is shown for different sizes of the threading window. The error bars indicate the standard deviation.

over the distance in sequence $|j-i|$ of the two residues i and j. Apparently the contact frequency varies only by a small amount for different pairs at the same distance in sequence. This simple dependence is expected from consideration of translation and inversion symmetry of the chain. That is to say, two residues should have the same properties, if they are shifted along the chain and if the sequence of residues is inverted. At the same time the translation symmetry is partly imprinted by the concept of threading, as one contact in a misfolded substructure contributes to pairs in the threading windows, that have the same distance in sequence.

As expected, the contact frequency decreases rapidly with the residue separation. At large distance this decrease can be very well described by a power law with a fitted exponent of -1.03. This exponent is smaller than the scaling of the return probability of a random walk, which is equal to -1.5, because misfolded structures are compact, and thus the chain is restricted small volume. Moreover, there are two interesting features of contact frequencies: First, at a distance of approximately 25 residues the contact frequency exhibits a bump, which was ascribed to the typical minimum length that a stiff chain needs to form a loop [43]. Second, at large inter residue distances the contact frequency slightly levels off, which can be understand from the properties of residues at the terminals of the chain. Since a pair residues which consists of two residues from both ends of the chain has a large sequence separation, it contributes much to the contact frequency at high $|j-i|$. The chain ends of native protein structures are known to be close in space [44], thereby contributing contacts to the average. This behavior, however, is is not expected for misfolded structures, which does not need a correction as the difference to the power law is rather small.

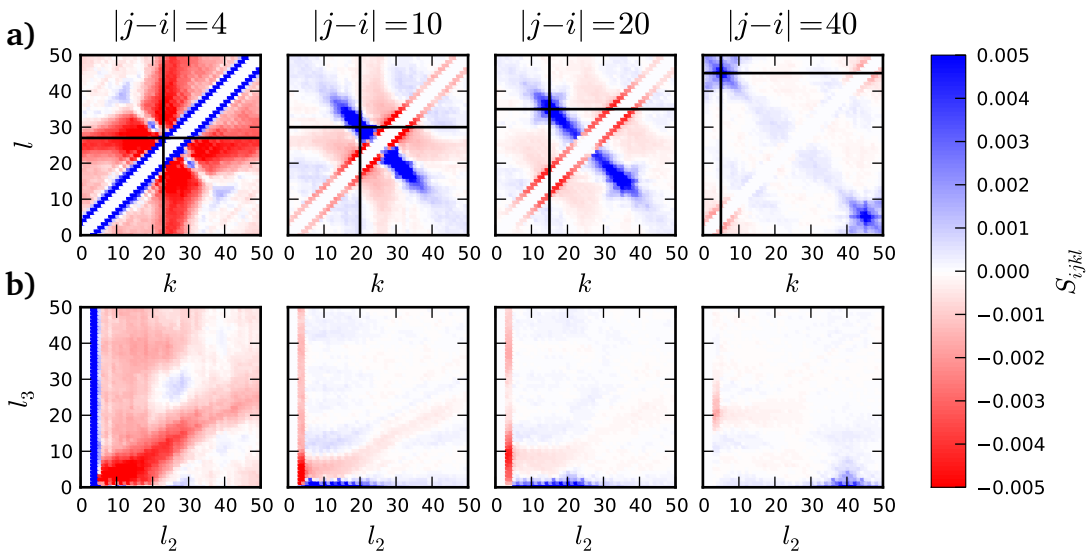

Figure 2.8: Contact correlations tensor S_{ijkl}. **a)** Measured by threading. The two crossing black lines indicate the pair (i, j). **b)** Measured in the homogeneous approximation. The homogeneous indices are $l_1 = |j - i|$, $l_2 = |l - k|$, and $l_3 = |(i + j) - (k + l)|/2$.

The contact correlation tensor S_{ijkl} consists of the two-contact-frequency $\left\langle C_{ij} C_{kl} \right\rangle$ and a product of the corresponding contact frequencies. Similar to the contact frequency, the two-contact-frequency can be measured by threading. For a threading window size of 50 residues the contact correlation tensor is depicted in Fig. 2.8.a). In general, one can discern a strong positive correlation of the contacts (k, l) and (i, j) if they are close in the contact matrix, i.e., if two residues from each pair are close in sequence. The correlation is particularly string, if the contacts share a contact. For the contact correlation of three contacts, which determines the third cumulant, the statistics of threading is too poor.

The correlation tensor reveals also correlations due to secondary structure elements. 90% of all short ranged contacts with $|j - i| = 3$ or $|j - i| = 4$ are helical contacts. A helical contact is strongly correlated with other helical contacts in the vicinity (left plot in Fig. 2.8.a)). That is to say, if a helix is formed the residues in the vicinity are likely to belong to the same helix. Since helices are very stiff, contacts between more distant residues in the helix cannot be formed and are therefore anti-correlated with helical contacts.

Correlations due to β-sheets are visible in Fig. 2.8.a) as stripes of positive correlations starting from the pair (i, j), which run parallel (parallel β-sheets with $|j - i| = |l - k|$) and orthogonal (anti-parallel β-sheets with $j + i = l + k$) to the diagonal.

The translation and inversion symmetry of the chain can be exploited to reduce the number of indices of the contact frequency and the contact correlation tensor. Thus, not the particular position of the residues i and j in a chain matters, but their relative position to each other, which gives raise to the two indices $l_1 = |j - i|$ and $l_2 = |k - l|$. The relative position of the two pairs of residues can be described by the index $l_3 = |(i + j) - (k + l)|/2$, which measures the distance of their centers. The division by two is an integer division, reflecting the fact that for certain values of l_1 and l_2 the amount $|(i + j) - (k + l)|$ can acquire only odd or even numbers. For illustration purposes the indices are shown schematically in Fig. 2.9.

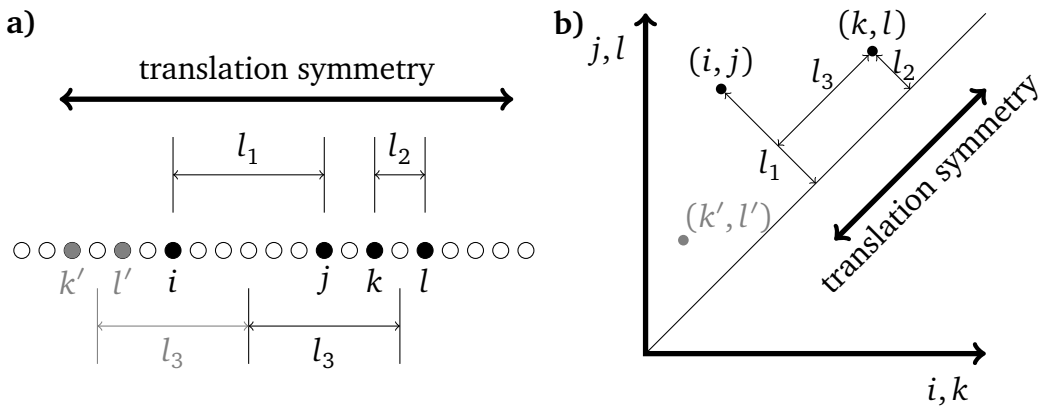

Figure 2.9: Illustration of the indices l_i used in the homogenous approximation in the chain **(a)** and the contact matrix **(b)**. The chain is symmetric under a shift along the chain (translation symmetry) and inversion of the sequence, that is, the pair (i,j) relates to the pair (k,l) in the same way as to the pair (k',l') (gray contacts/residues).

Along with the symmetries comes the homogeneous approximation, which assumes that the correlation tensor depends only on the three indices l_1, l_2, and l_3. This approximation can be tested by comparing the entries of the correlation tensor measured by threading that belong to the same homogeneous indices. The standard deviation of the correlation tensor entries is small compared to the mean and therefore the homogeneous approximation is justified (see Fig. A.2).

To determine the contact frequency in the homogeneous approximation, I determine the number of possible contacts $\mathrm{PC}(l,L_p)$ for each protein p of the threading set, i.e., the number of residue pairs, in a sequence separation l. These depend only on the length L_p of the protein, and the number of contacts $\mathrm{AC}(l,p)$ in sequence separation l. These two quantities are summed over all proteins and the ratio defines the contact frequency $w_C(l)$,

$$w_C(l) = \frac{\sum_p \mathrm{AC}(l,p)}{\sum_p \mathrm{PC}(l,L_p)} \quad . \tag{2.14}$$

Likewise, to measure the two-contact-frequency in the homogeneous approximation, the number of Possible Contact Pairs $\mathrm{PCP}(l_1,l_2,l_3,L_p)$ and the number of Actual Contact Pairs $\mathrm{ACP}(l_1,l_2,l_3,p)$ are introduced. The ACP measures the number of residue pairs, where both pairs are in contact. By summing the two quantities over all proteins in the threading set, one finds the estimate w_{CC} for the two-contact-frequency $\left\langle C_{ij}C_{kl}\right\rangle$,

$$w_{CC}(l_1,l_2,l_3) = \frac{\sum_p \mathrm{ACP}(l_1,l_2,l_3,p)}{\sum_p \mathrm{PCP}(l_1,l_2,l_3,L_p)} \quad . \tag{2.15}$$

While the contact correlations measured by threading automatically fulfill the normalization conditions eqs. (2.13), the homogeneous approximation is not correctly normalized. Indeed, the contact correlation tensor in the homogeneous correlations normalizes to a strongly negative number (see Fig. 2.10.a)). To avoid such inconsistencies, contact frequency and contact correlation are multiplied by a normalization factor such that their normalization is identical to the fitted moments of contact correlation measured by threading (cf. eqs. (2.3)).

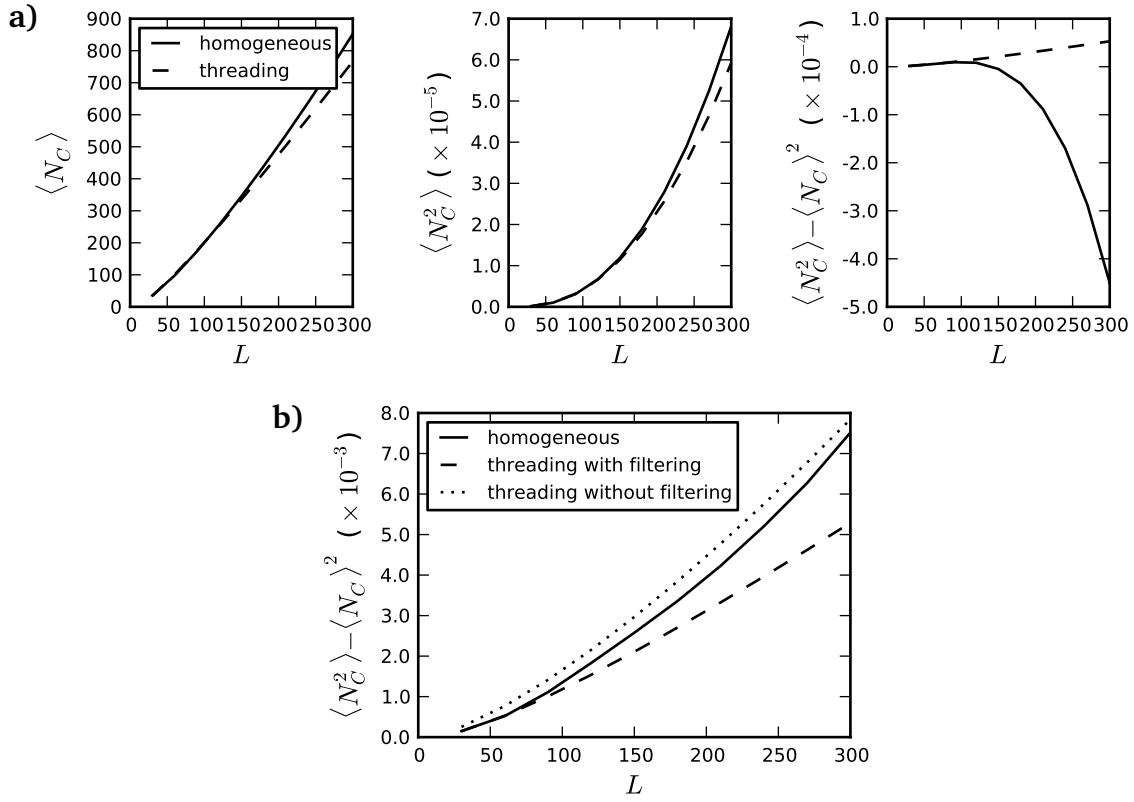

Figure 2.10: Normalization of homogeneous approximation. **a)** Contact frequency with one index, **b)** contact frequency with three indices.

Thus, the contact frequency $\left\langle C_{ij} \right\rangle_L$ and the two-contact frequency $\left\langle C_{ij}\, C_{kl} \right\rangle_L$ read the homogeneous approximation like

$$\left\langle C_{ij} \right\rangle_L = \langle N_C \rangle_L \, \frac{w_C(l_1)}{\sum_{l_1'} w_C(l_1') \mathrm{PC}(l_1', L)} \tag{2.16}$$

$$\left\langle C_{ij}\, C_{kl} \right\rangle_L = \left\langle N_C^2 \right\rangle_L \, \frac{w_{CC}(l_1, l_2, l_3)}{\sum_{l_1', l_2', l_3'} w_{CC}(l_1', l_2', l_3') \mathrm{PCP}(l_1', l_2', l_3', L)} \,. \tag{2.17}$$

In this way, the following normalization condition are fulfilled

$$\sum_{l_1} \left\langle C_{ij} \right\rangle_L \mathrm{PC}(l_1, L) = \langle N_C \rangle_L \tag{2.18}$$

$$\sum_{l_1, l_2, l_3} \left\langle C_{ij}\, C_{kl} \right\rangle_L \mathrm{PCP}(l_1, l_2, l_3, L) = \left\langle N_C^2 \right\rangle_L \,. \tag{2.19}$$

However, this normalization of the contact correlations is only a provisional. The deeper reason for the mismatch of $w_C(l)$ and $w_{CC}(l_1, l_2, l_3)$ is that both depend on different indices. By extending the definition of the contact frequency to a quantity that depends on the three indices $l1$, $l2$, and $l3$, the problem of the correct normalization of the contact correlation S_{ijkl} can be circumvented.

The idea behind the new definition is that the contact frequency for a pair (i,j) can be computed by summing over the pair probability $P(C_{ij}=1, C_{kl}=C')$,

$$\left\langle C_{ij}\right\rangle_{kl} \equiv P_{kl}(C_{ij}=1) = \sum_{C'=0}^{1} P(C_{ij}=1, C_{kl}=C') = \left\langle C_{ij}C_{kl}\right\rangle + \left\langle C_{ij}(1-C_{kl})\right\rangle \quad . \tag{2.20}$$

Note that for threading $P_{kl}(C_{ij}=1)$ is the same for all (k,l) by definition and is equal to $\left\langle C_{ij}\right\rangle_{\text{threading}}$.

To this end, the variable $\text{NCP}_1(l_1,l_2,l_3,p)$ is defined, which measures the number of pairs of residues pairs, where the pair that corresponds to l_1 is in contact. Similarly $\text{NCP}_2(l_1,l_2,l_3,p)$ is identical to the number of four residues (i,j,k,l), where the l_2 pair is in contact. Again, I adopt the homogeneous approximation and introduce two contact frequencies $\langle C_1\rangle(l_1,l_2,l_3)$ and $\langle C_2\rangle(l_1,l_2,l_3)$ that should match better to a pair of residue pairs with indices l_1, l_2 and l_3 than the contact frequencies $w_C(l_1)$ and $w_C(l_2)$. $\langle C_1\rangle(l_1,l_2,l_3)$ corresponds to the contact frequency of the contacts with inter-residue distance l_1 and $\langle C_2\rangle(l_1,l_2,l_3)$ to inter-residue distance l_2. Thus $\langle C_1\rangle(l_1,l_2,l_3)$ is expected to be similar to $w_C(l_1)$ and $\langle C_2\rangle(l_1,l_2,l_3)$ to $w_C(l_2)$. The new contact frequencies are defined as,

$$\langle C_i\rangle(l_1,l_2,l_3) = \frac{\sum_p \text{NCP}_i(l_1,l_2,l_3,p)}{\sum_p \text{PCP}(l_1,l_2,l_2,L_p)} \quad . \tag{2.21}$$

The contact correlation yields a normalization that is close to the normalization observed in threading without the need of a normalization. Therefore, for computations below, the three index version of the contact correlation tensor is adopted. Although it is not needed, the contact correlation is multiplied with a normalization factor, which is not very different from one (see Fig. 2.10.b)). The homogeneous approximation does not filter non-compact structures as in threading. Accordingly, the normalization of the contact correlation tensor is closer to the contact variance of unfiltered threading.

2.4.2 Selection on free energy

Contact correlations have a considerable contribution to cumulants and therefore to the misfolded free energy. The question is, whether one can identify selection in data of wild type proteins that show selection that explicitly uses contact correlations. To this end, I compare the free energy of wild type sequences to two kinds of randomized sequences. First, I will compare to random sequences, i.e., every amino acid is drawn independently from the distribution of amino acids observed in the PDB (see Table A.3), to test for selection of amino acid composition and positional correlation. Second, I will compare to shuffled sequences, which have the very same amino acid composition of the corresponding wild type sequence. Here, only the positional correlation in sequence is relevant.

However, the comparison to randomized sequences comes with a caveat. The test works in favor of the randomized sequences as wild type sequences, in contrast to randomized sequences, are subject to selection of positive and negative design. Consequently, wild type sequences have to have a low free energy in the native state, while randomized sequences are not subject to such a constraint. In fact, the free energy of the native state and misfolded ensemble are

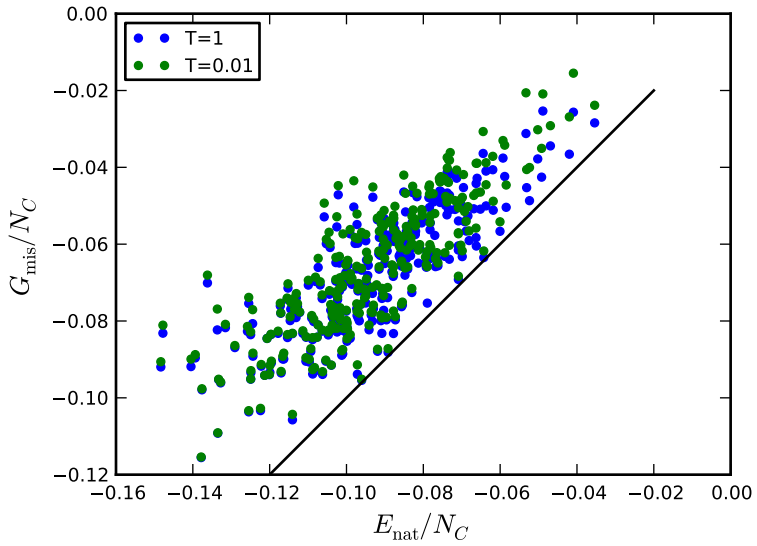

Figure 2.11: Native energy and misfolded free energy are correlated. Black line indicates line of stability at which $\Delta G = 0$ holds.

strongly correlated (see Fig. 2.11). Thus, sequences with a lower misfolded free energy are not necessarily more prone to misfolding as they have a comparable free energy difference ΔG to sequences that have a higher misfolding free energy.

For the test, I define a test set of 300 chains that are a random subset of the rank one structures of the PDB sequence cluster with 50% sequence identity. To define the properties of the test set, chains are restricted to a length of 50 to 300 residues and are removed from the set if they were non-compact, i.e., if their number of contacts N_C is less than one standard deviation below the mean number of contacts expected for the specific length ($N_C/N > 3.6 - 7.5\,L^{-}1/3$). If a protein consists of more than on chain, the stability of the structure of one chain might depend on the interaction with other chains, which might yield spurious signals. Therefore, chains that have many contacts with other chains, i.e., if the ratio of inter to intra chain contacts is larger than 0.15, are removed from the set. Structures which were not determined with X-ray crystallography, which is very accurate, are removed from the set. To avoid membrane proteins, wild type sequences with a large hydrophobicity $[h] > 0.17$ are removed. If the wild type sequence from the PDB file contained any undetermined amino acids, it is rejected.

For each sequence I generate 100 random sequences and 100 shuffled sequences, for which I measure the cumulants by threading and compute the free energy using the expansion up to a certain order. The free energy of the wild type sequence is compared to the distribution of randomized sequences. One measure is the fraction $P_<$ of randomized sequences that have a misfolded free energy below the free energy of the wild type sequence. A measure, which is widely used, is the Z-score of the wild type sequence with respect to the distribution of the randomized sequences, which is highly correlated with $P_<$,

$$\text{Z-score} = \frac{x_{\text{wild type}} - \langle x_{\text{rand}} \rangle}{\sigma(x_{\text{rand}})} \quad . \tag{2.22}$$

Due to the problem of the wrong entropy of the free energy, the free energy is computed from the expansion, using cumulants that are determined from threading. Furthermore, this

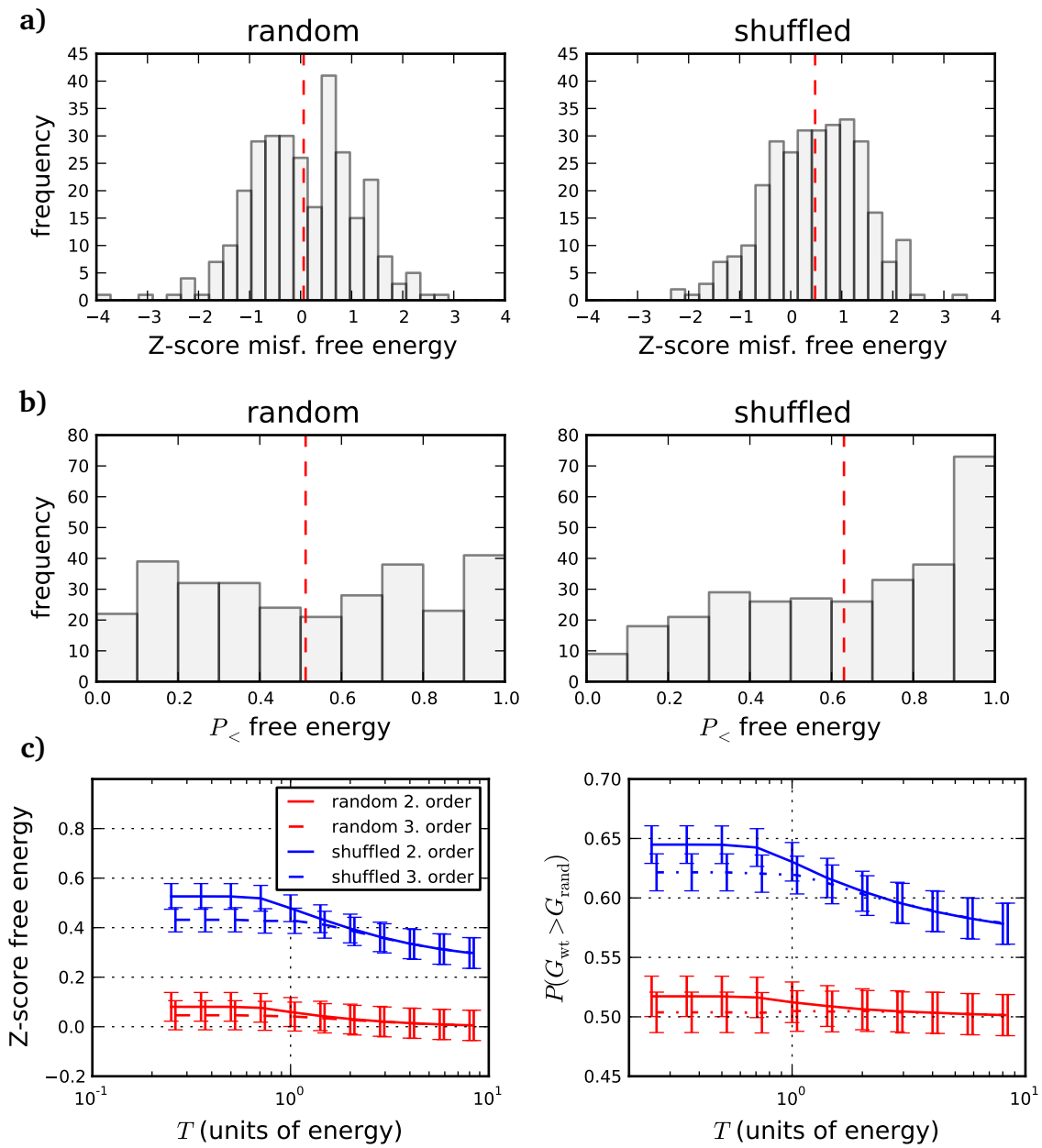

Figure 2.12: Average Z-score of free energy versus temperature. The Z-score is computed with respect to random and shuffled sequences. The free energy is computed from the expansion up to second and third order.

approach allows for subdividing the influences of the cumulants to the free energy. For the temperature that is considered relevant, i.e., $T = 1.2$ energy units, the Fig. 2.12 shows the distribution of the two measures for the test set is shown for a free energy at temperature $T = 1.2$. The free energy was computed from the second order expansion.

If wild type sequences did not differ from randomized sequences, the distribution of the Z-score would be a normal distribution and the distribution of the rank would be a uniform distribution. Indeed, the distributions of the measures vary little from this expectation, that is, wild type sequences have a similar misfolded free energy to random and shuffled sequences. Nevertheless, one can discern significant deviations from the random expectation. The average

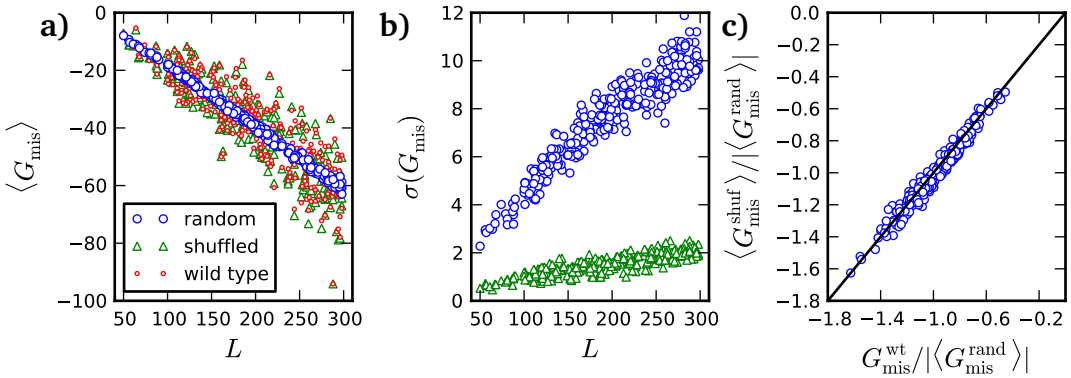

Figure 2.13: Length scaling of free energy of randomized and wild type sequences at $T = 1.2$. Misfolded free energy versus length for wild type sequences the mean **a)** and standard deviation **b)** of randomized sequences. **c)** Comparison of wild type with the mean of shuffled sequences, the length scaling is removed by dividing by the mean for random sequences.

Z-score and average $P_<$ is larger than zero, indicating that wild type sequences have a larger free energy than randomized sequences.

The signal is much stronger for shuffled sequences than for random sequences. Random sequences cover a large spectrum ranging from very hydrophobic to very hydrophilic sequences. This range is covered also by wild type sequences, which therefore seem to be indistinguishable from random sequences. Shuffled sequences, on the other hand, have the same sequence composition and their misfolded free energy is narrowly distributed about the one of wild type sequences (see Fig. 2.13). Consequently, it is remarkable that wild type sequences have a significantly larger misfolded free energy than the respective shuffled ensemble.

In Fig. 2.12.c) the temperature dependence of the two scores is shown. At small temperatures the free energy is identical with the free energy at freezing temperature, where the first three cumulants have a major contribution. With increasing temperature the free energy is dominated by the first cumulant and the entropy, which, however, does not depend on sequence composition. That is, with increasing temperature the two selection measures approach the corresponding values of the first cumulant.

The selection measures decrease with temperature and the selection measure of the third order expansion is smaller than for the second order. This contribution of the individual cumulants will be investigated in the next Section.

2.4.3 Energy cumulants

In this Section, the selection on individual energy cumulants it is investigated by comparing cumulants of wild type sequences to cumulants of random and shuffled sequences. The scaling with length of the mean and standard deviation of the cumulants for random and shuffled sequences are shown in Fig. 2.14. The mean of the cumulants of random sequences can well be fitted by $\epsilon_n \langle N_C \rangle_L$ (see Fig. 2.14.a)). The values for ϵ_i are listed in Table 2.2 and are compared to the mean interaction energies, which result from the composition of random sequences. The standard deviation of random sequences is always larger than for shuffled sequences, however,

	from fit	from amino acid frequency	
ϵ_1	-0.0139 ± 0.00008	$[U]$	-0.0155
ϵ_2	0.0403 ± 0.00005	$[U^2]$	0.0335
ϵ_3	-0.0206 ± 0.00008	$[U^3]$	-0.00868
ϵ_4	0.0370 ± 0.00025	$[U^4]$	0.00499

Table 2.2: Fitted parameters ϵ_n and moments from energy.

the standard deviation of shuffled sequences becomes more similar to the standard deviation of random sequences with increasing order of the cumulants. Since higher cumulants are sensitive to small changes in the tails of a distribution, their variation increases for shuffled sequences.

The normalized mean of the cumulants is close to the value of wild type sequences, however, the correlation decreases with the order of the cumulant. That is, the first and second cumulants for shuffled sequences are narrowly distributed about the wild type sequence, while random sequences cover the broad range of values, that arise from different sequence compositions.

Similar to the free energy, the selection measures $P_<$ and the Z-score do not show a significant difference between wild type and random sequences (data not shown) and the selection becomes visible only in the comparison to shuffled sequences (see Fig. 2.15). The first cumulant only has a small difference to shuffled sequences. The distribution of $P_>$ shows a marginally significant increase of value at the upper end of the range for wild type sequences. The average Z-score is 0.26, which is significantly larger than zero. What is interesting, is that the Z-score of the first cumulant is significantly negatively correlated with the number of helical, i.e., short range, contacts (data not shown). Proteins with many helices can increase their stability against unfolding if they make interactions of short range contacts more attractive. However, attractive short range contacts decrease also the first cumulant of the misfolded ensemble. By shuffling the sequence, these short range sequence correlations are broken apart and the shuffled ensemble can acquire a larger first cumulant. Thus, if there was a way to disentangle the contributions of positive and negative design, it would be reasonable to assume that the signal of negative design improves.

The second cumulant, however, shows the strongest signal, showing that large positive second cumulants are much less often in wild type sequences than in random sequences. The third cumulant is more tricky. It shows that values at the positive tail of the distribution are suppressed, at which the third cumulant makes a contribution that suites the misfolding stability.

The fourth cumulant shows clear signal that is not in accordance with negative design. Cumulants are heavily influenced by short range contacts. Tests with pairs with less than four diagonals showed, that the signal of the fourth cumulant is inverted (data not shown).

After all, two problems make the results difficult to interpret. First, negative and positive design shape wild type sequences and properly disentangling their contributions is difficult. Second, cumulants are strongly correlated and a selection on one cumulant, in order to increase the stability against misfolding, can change another cumulant that yields the opposite signal.

Selection is expected to act on extremely unfavorable values. Therefore, the extreme values of wild type and randomized sequences are compared to each other. To remove the length dependence of the cumulants, they are transformed to Z-scores with respect to the random

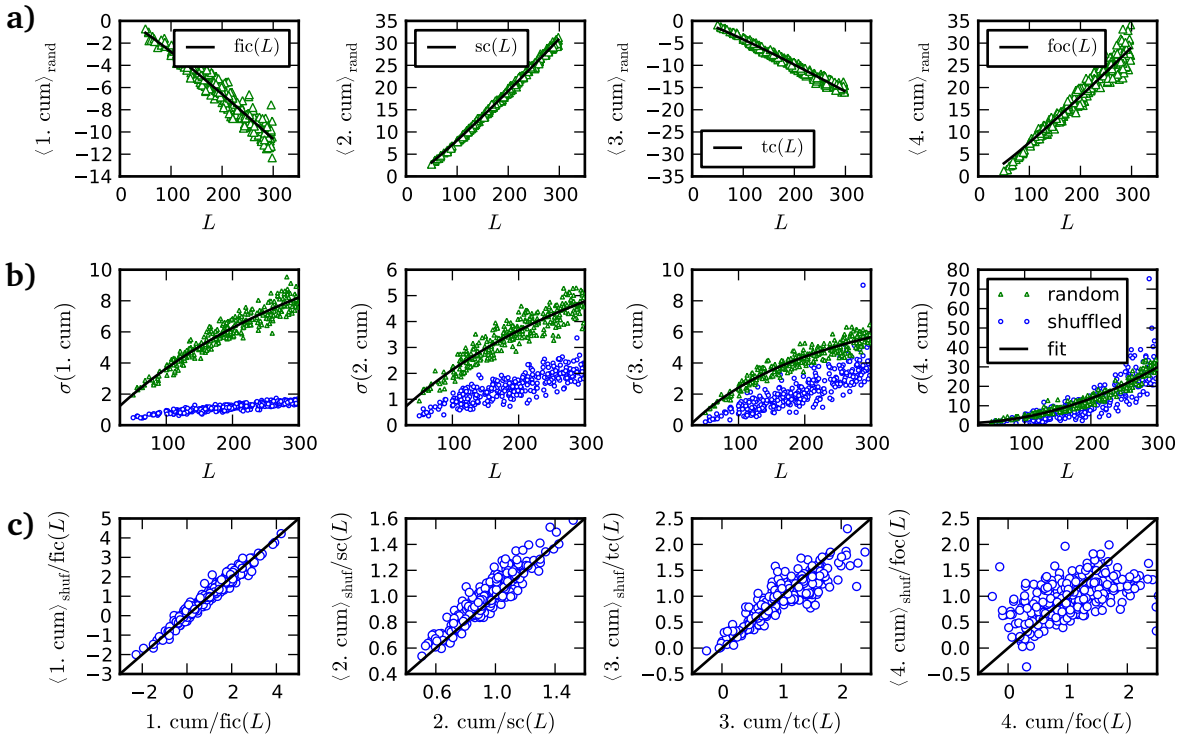

Figure 2.14: Length scaling of free energy cumulants. **a)** The average cumulant of random sequences can be fitted very well by $\epsilon_i \langle N_C \rangle_L$, **b)** standard deviation of cumulant of random and shuffled sequences versus length, **c)** comparison of the normalized mean of shuffled sequences to wild type sequences.

ensemble, where the fitted mean and standard deviation as function of length are used. The standard deviation is fitted by the functions (see Fig. 2.14.b))

$$\text{stdDev}(\text{first cumulant})(L) = 21.884 \frac{L}{L + 499.7} \tag{2.23a}$$

$$\text{stdDev}(\text{second cumulant})(L) = 12.49 \frac{L}{L + 484.3} \tag{2.23b}$$

$$\text{stdDev}(\text{third cumulant})(L) = \frac{(L - 26.88)\, 11.16}{L + 238.6} \tag{2.23c}$$

$$\text{stdDev}(\text{fourth cumulant})(L) = L^2\, 0.0003191 + 1 \quad . \tag{2.23d}$$

Then, the number of cumulants above or below a threshold is evaluated. The threshold is such that 5% of random sequences are beneath it. The first and third cumulants wild type sequences are not different from randomized sequences, but the second and fourth cumulants produce a significant signal (see Fig. 2.16). For the second cumulant it is in favor for negative design, whereas it is not for the fourth cumulant. Theses results are in accordance with the observations of the analysis of $P_<$ and the Z-score. Again, one can argue that the fourth cumulant is correlated with other cumulants.

In addition to the cumulants, Fig. 2.16 shows selection for avoiding extreme values of the hydrophobicity. The value for shuffled and wild type sequences are equal by definition. The

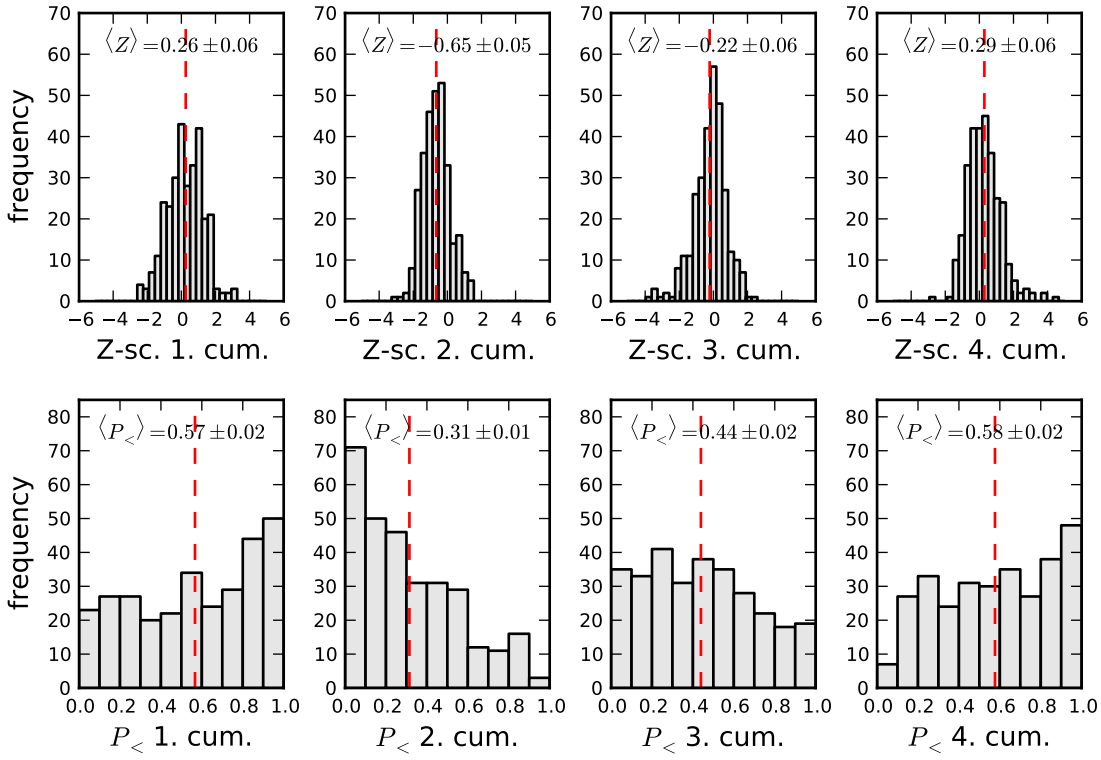

Figure 2.15: Z-score and $P(\text{shuffled} < \text{wt})$ of cumulants with respect to shuffled sequences.

comparison with random sequences shows that extremely hydrophobic wild type sequences are suppressed, in accordance with the observation that the content of extremely hydrophobic and extremely hydrophilic amino acids are correlated (cf. Section 2.3.4).

2.4.4 Negative design scores

It was shown in the last Section that the first and second cumulants of wild type sequences carry signals of selection of negative design. In this Section scores are defined that assess the amount of selection.

Selection of the first cumulant can be detected if the average contact frequency $\left\langle C_{ij} \right\rangle$ is positively correlated with the interaction energy. Thus, the most simple and practical score, called contact frequency energy score (CFES), is the correlation coefficient between contact frequency and interaction energy,

$$\text{CFES} = \text{Corr.coeff.}\left(\left\langle C_{ij} \right\rangle, U_{ij}\right) \quad , \tag{2.24}$$

where pairs with $|j - i| < 3$ are omitted. The advantage of the correlation coefficient is that it is restricted to an interval, ranging from -1 to 1, and has no bias that arise from rescaling or shifting of variables. The score is computed from the contact frequency measured in the homogeneous approximation according to eq. (2.16).

To assess the question to what extend contact correlations are exploited by negative design, it is useful to decompose the second cumulant into terms that describe different levels of cor-

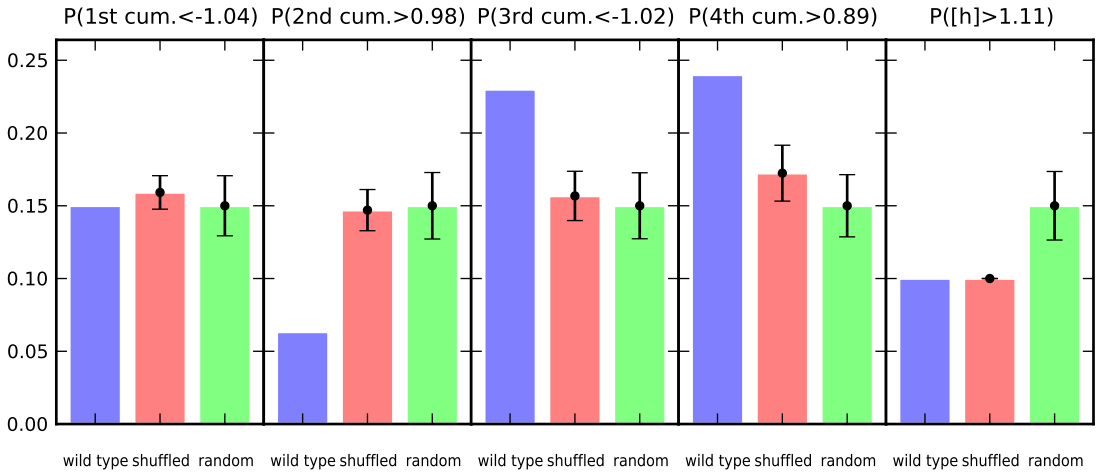

Figure 2.16: Extreme values of cumulants and hydrophobicity. Bars indicate the probability that the Z-score of energy cumulant exceeds the threshold indicated above the plot. The fifth plot refers to the distribution of hydrophobicity.

relation. By substituting the product $U_{ij}U_{kl}$ by $(U_{ij} - [U])(U_{kl} - [U])$ in the definition of the second cumulant, one finds

$$\sum_{i<j-2,k<l-2} S_{ijkl}\, U_{ij}\, U_{kl} = \left\langle (N_C - \langle N_C \rangle)^2 \right\rangle [U]^2 + 2\,[U] \sum_{ij} D_{ij} \left(U_{ij} - [U]\right) + \sum_{i<j-2,k<l-2} S_{ijkl} \left(U_{ij} - [U]\right) \left(U_{kl} - [U]\right) \quad , \tag{2.25}$$

where the term $D_{ij} = \sum_{k<l-2} S_{ijkl}$ denotes the overall correlations of the residue pair (i,j) with all other pairs. The effect of sequence composition is represented by the term $\left\langle (N_C - \langle N_C \rangle)^2 \right\rangle [U]^2$ and the contribution of contact correlation are comprised in the summands multiplied with S_{ijkl}. In Fig. 2.17.a) the contribution of the three terms to the second cumulant is depicted. Since $[U]$ is very small, the D_{ij} term and $\left\langle (N_C - \langle N_C \rangle)^2 \right\rangle [U]^2$ are very small, where the former can easily be neglected. The latter depends solely on sequence composition and is identical for wild type and shuffled sequences.

Based on this decomposition, the contact correlation energy score (CCPES) is defined as the correlation coefficient of the product of interaction energies and the contact correlation tensor from the homogeneous approximation,

$$\text{CCPES} = -\text{Corr.coeff.}\left(S_{ijkl}, \left(U_{ij} - [U]\right)\left(U_{kl} - [U]\right)\right) \quad . \tag{2.26}$$

If either of the pairs (i,j) or (k,l) is closer in sequence than three residues, S_{ijkl} is zero and does not contribute to the second cumulant. Consequently, the corresponding terms are omitted from the computation of the score. The CCPE score, however, suffers from a bias, that becomes discernible if the score is measured for shuffled sequences. The bias stems from two sources. First, if the two pairs (i,j) and (k,l) are identical, the interaction term $(U_{ij} - [U])^2$ as well as the contact correlation $S_{ijij} = \langle C_{ij} \rangle - \langle C_{ij} \rangle^2$ are positive, but all other pairs with $(i,j) \neq (k,l)$ can acquire both signs. This difference in signs causes a bias. Omitting identical pairs, however, does not remove the bias.

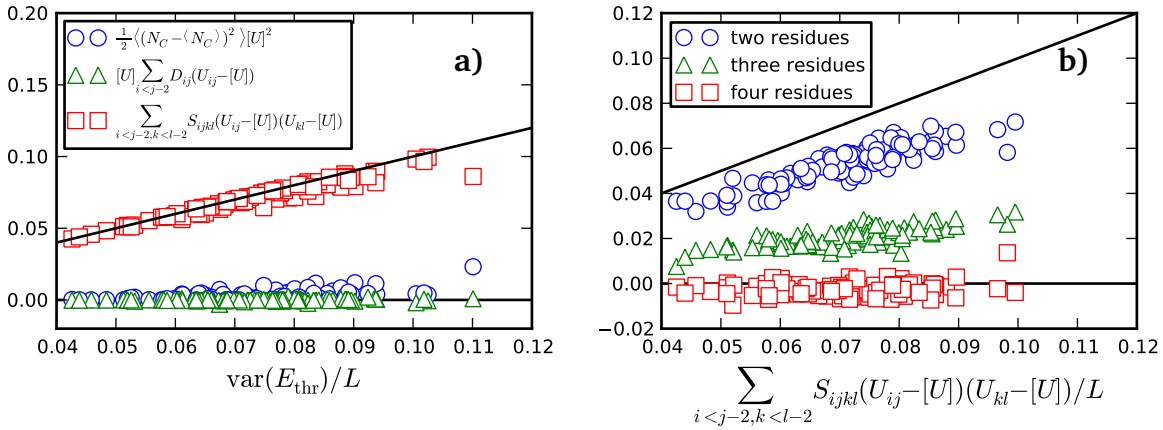

Figure 2.17: Contribution to second cumulant of misfolded free energy. **a)** mean field separation. **b)** contribution to correlation term: Most of the variance is contributed by the correlation of the contact with itself.

Second, the correlation coefficient considers residue pair quantities that are shaped by properties of single sites (for instance hydrophobicity). Pairs which share a side are therefore correlated. Interestingly, these biases vanish if the score is split into three scores, which measure the correlation of terms that have two, three or four different residues and are referred to as CCPES2, CCEPS3, and CCPES4 respectively. Furthermore, the splitting allows to asses the difference in selection on these terms. In fact, one might expect larger scores for terms with less different residues, as less residues are subjected to fewer, possibly competing constraints. This is consistent with the observation, that the contribution to the second cumulant of the terms decreases with increasing number of different residues (see Fig. 2.17.b).

To disentangle positive and negative design, all scores are evaluated for native or non-native contacts separately, giving rise to variants of the scores defined above, whose names are appended “nat” (score for native contacts) and “nonat” (score for non-native contacts) in the following. As the three and four residue variant of the CCPE score comprise two different pairs, it is principally possible to consider a version of the score, which considers a native and a non-native contact. However, since this would increase the number scores to consider, these variants are neglected for clarity.

The correlation coefficients, which constitute the scores, are small and attain a broad range of positive and negative values (see Fig. 2.18). As before, the signal of negative design is seen best, if averages over many sequences are considered. In general, the average of most scores for wild type sequences are significantly above zero, while they are indistinguishably from zero for shuffled sequences, as expected for an unbiased score, showing that wild type sequences are shaped by negative design. This signal is even more pronounced if the scores are computed for native and non-native contacts, where the distribution is significantly shifted to more positive values. Interestingly, the shift is larger for the native contacts, probably because the set of native contacts is smaller than the set of non-contacts.

Generally, the more different residues are considered the magnitude of the CCPE scores decreases largely. This is because the scores, which consider more different residues, consider more terms, where many of these terms have a S_{ijkl} that is close to zero. Nevertheless, the

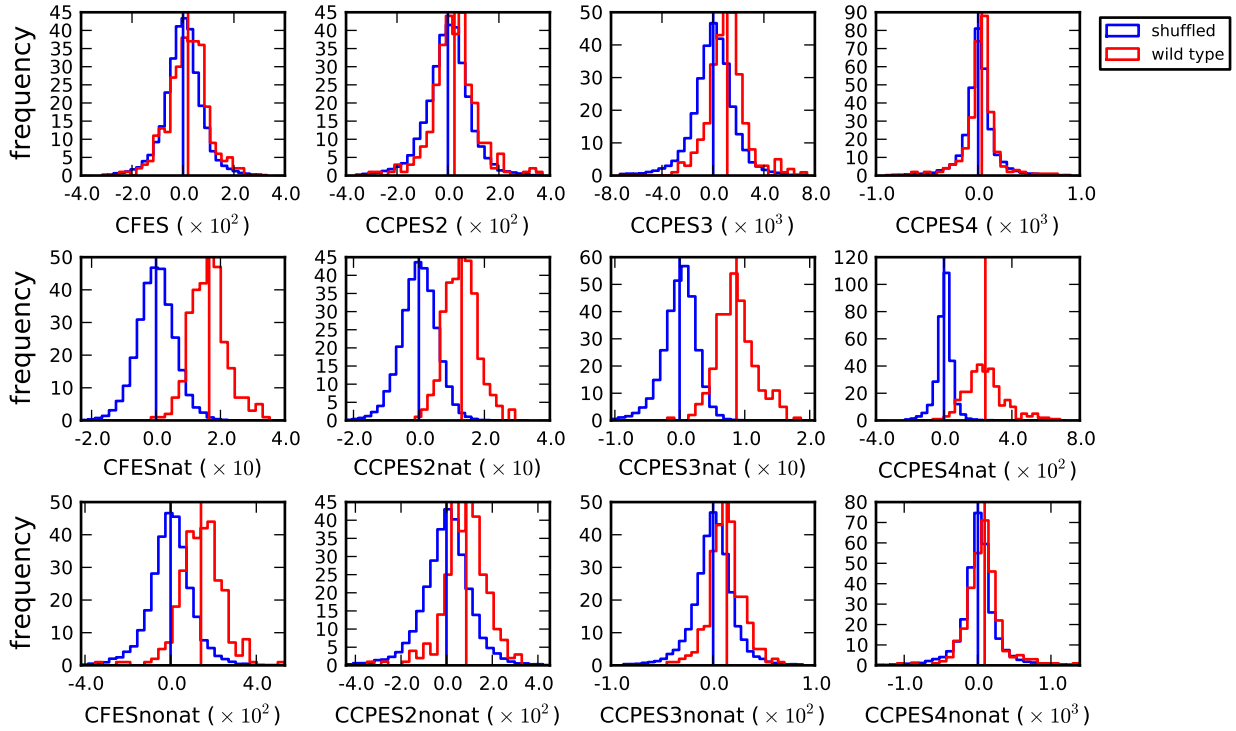

Figure 2.18: Histogram of negative design scores for wild type and shuffled sequences. Horizontal lines indicate the mean of the distribution.

difference in magnitude reflects the expectations and the decreasing contribution to the second cumulant (fig 2.17.b)).

To assess the significance of the score, I compute for each wild type sequence the Z-score with respect to the distribution of 50 shuffled sequences (histogram in Fig. 2.19 on page 38). In accordance with the weak selection of the first cumulant, the mean CFE score is only marginally above zero. Remarkably, the score for native (CFESnat) and non-native (CFESnonat) contacts is much larger and highly significant. This can be understood by considering the difference ΔG of native and misfolded free energy estimated by the first cumulant,

$$\Delta G \approx \sum_{i<j-2} (C_{ij}^{\text{nat}} - \langle C_{ij} \rangle) U_{ij} \quad .$$

Thus, the free energy difference is decreased if the interaction energy is negatively correlated with $C_{ij}^{\text{nat}} - \langle C_{ij} \rangle$. Fig. 2.20 shows the mean interaction energy of the test set binned by the variable $C_{ij}^{\text{nat}} - \langle C_{ij} \rangle$. Clearly, the average interaction energy decreases with this variable. Since the contact frequency depends on sequence separation, the average interaction energy decreases with sequence separation (see Fig. 2.20.a) and inset in Fig. 2.20.b)). If all pairs are considered, i.e., native and non-native contacts are combined, the decrease with sequence separation is much weaker (see Fig. 2.20.a)), even though native and non-native contacts exhibit a clear decrease. This can be understood by the fact that native contacts contribute more to the average of all contacts at small sequence separation than at large separation.

Of particular interest is the dependence on $\left[U\right]$ and on the length of the chain. The expectation is, that sequences with large negative $\left[U\right]$ are more optimized with respect to negative

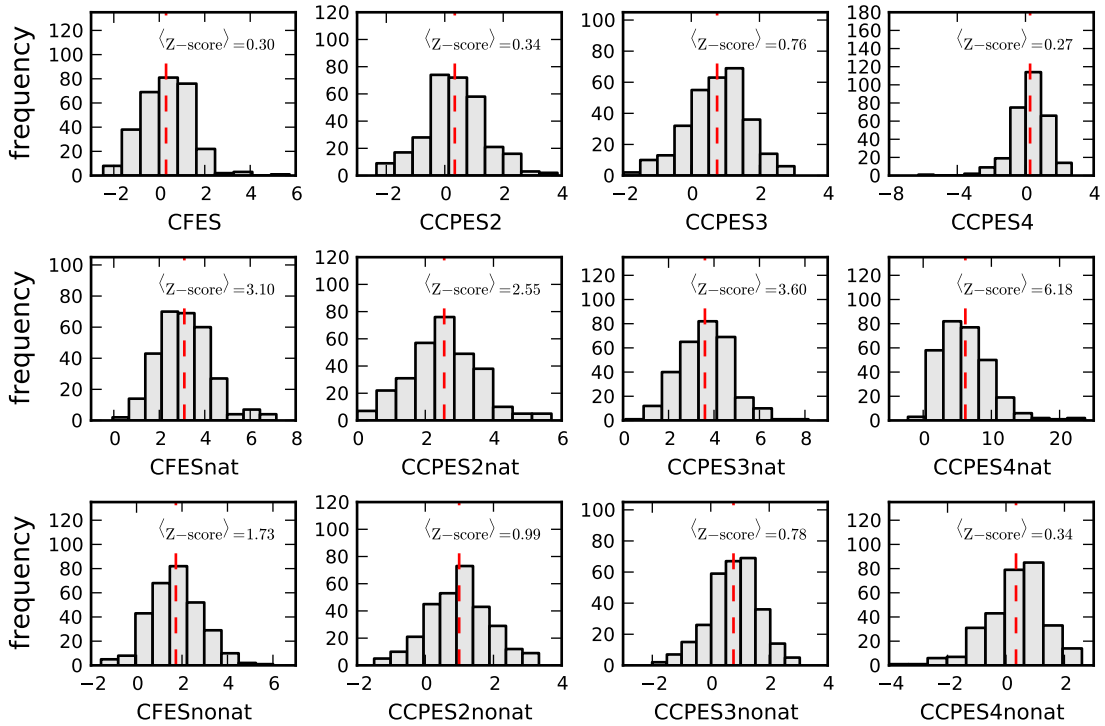

Figure 2.19: Distribution of Z-scores of negative design scores. Red lines indicate the mean of the distribution.

design, even though it was shown that sequences with a lower $\lceil U \rceil$ are also more stable against unfolding. Longer chains are expected to be better designed against misfolding, as they are more prone to get trapped in a misfolded structure.

The raw scores and Z-scores are binned into five bins of $\lceil U \rceil$ or length, each of which contains 60 proteins. The dependence on $\lceil U \rceil$ is very similar for the mean raw scores (see Fig. 2.21.a) on page 40) and Z-scores (see Fig. 2.22.a)). The CFE score does not show a correlation with the mean interaction energy $\lceil U \rceil$. The corresponding scores for the native contacts, however, show a significant decrease, while the decrease is only discernible at a very large $\lceil U \rceil$ for non-native contacts.

In general, the scores decrease with length due to length scaling of the contact frequency and the contact correlation (see Fig. 2.21.b)), which are effectively zero for large sequence separations. Therefore, it is important to assess the length dependence in terms of the Z-score with respect to shuffled sequences (see Fig. 2.22.b)).

In contrast to the raw scores, the Z-scores of CCPES3 and CCPES4 show an increase with length, while the CCPES2 and CFES score acquire at least a positive value throughout all lengths. The Z-scores of the variants for native and non-native contacts show a clear increase with length, with exception of the CCPES4nonat score, which is very small anyway.

By shuffling the sequence, the set of native contacts consists of different amino acid pairs for wild type and shuffled sequences. To ensure that the signal of the Z-score is not caused by this difference, I computed the Z-score, where the exact same amino acids where assigned to the native and non-native contact class for wild type and shuffled sequences. In other words,

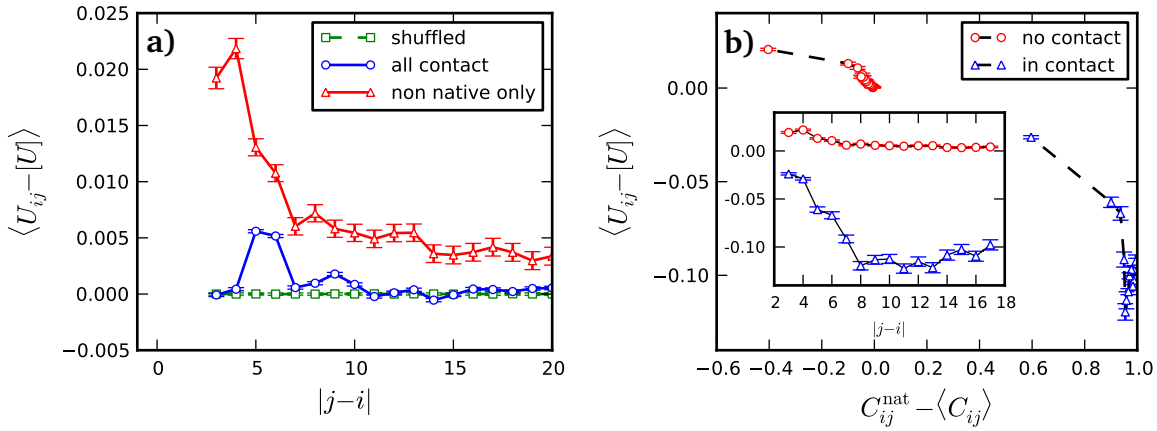

Figure 2.20: Interaction energy binned by (non-)native contacts. **a)** The deviation of interaction energy U_{ij} from the mean $[U]$ averaged in bins of sequences separation $|j-i|$, even though most pairs are non-native contacts the decrease in interaction energy is stronger if the native contacts are excluded. **b)** The deviation from the mean is binned in bins of $C_{ij}^{\mathrm{nat}} - \langle C_{ij} \rangle$. The decrease with contact frequency is reflected in the decrease with sequences separation (inset).

the contact matrix was shuffled in the same way as the sequence. The Z-score did not change significantly (data not shown).

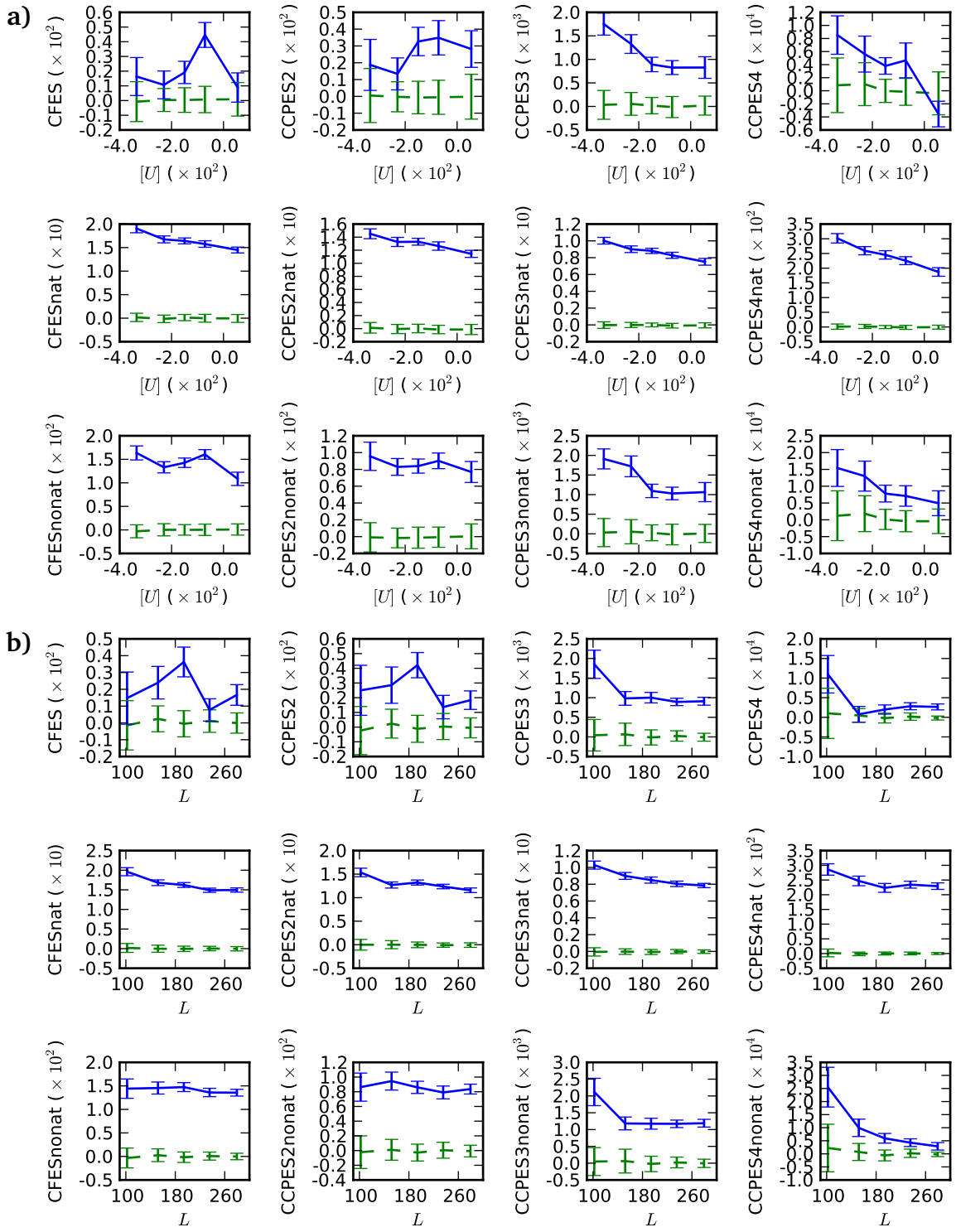

Figure 2.21: Negative design scores binned by $[U]$ (**a**) and length (**b**). The scores for shuffled sequences are on average zero. The errorbars for real sequences are estimated from the standard deviation in each bin. For shuffled sequences they are determined from the standard deviation of 50 shuffled samples. Note that the ordinate data is shown in different magnitudes.

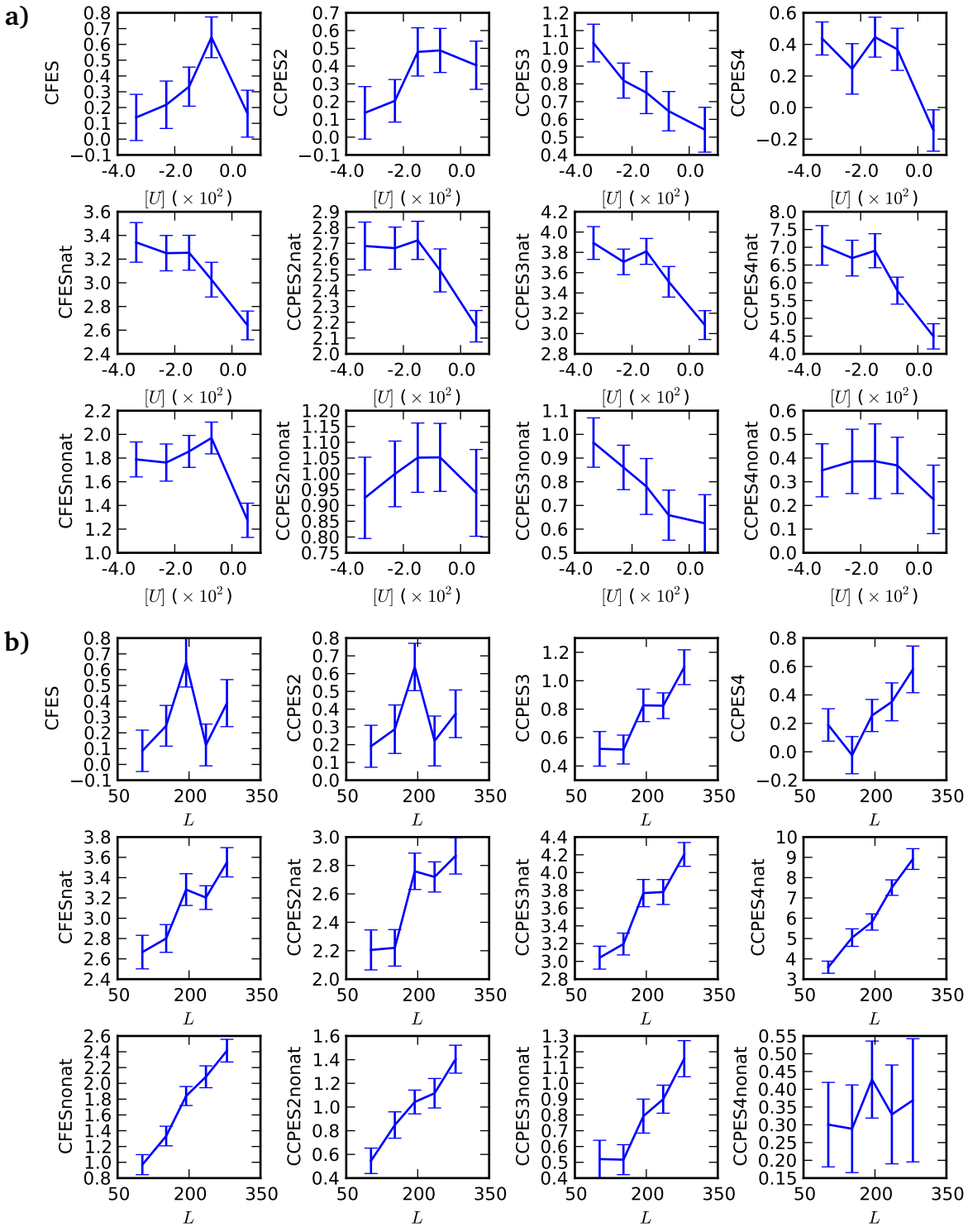

Figure 2.22: Average Z-scores of negative design scores binned by $[U]$ (**a**) and length (**b**). Error bars indicate the error of the mean.

2.4.5 Optimal hydrophobicity profiles

For the design of new proteins or proteins with new sequence properties, it is interesting to design sequences that fold into a specific three dimensional structure. As laid out in the introductory chapter, positive and negative design have to be considered. Indeed there exists studies which explicitly take into account negative design. These studies maximize the probability of the native state in a Boltzmann ensemble consisting of the desired native state and misfolded structures. This strategy is equivalent to maximizing the free energy difference ΔG. To find the optimal sequence, Monte Carlo simulations where employed, where single amino acids where mutated, to find the optimal sequence, starting from a random sequence. In two of these studies the free energy of the misfolded ensemble was estimated by sampling the structure space [45, 46], while the study of Morrissey *et al.* used a cumulant expansion, where, in contrast to this work, the cumulants were computed using the independent contact approximation [37]. The cumulant expansion allowed them to designed sequences with optimal stability for a specific temperature. Jin *et al.* designed the sequence of a small protein, whose native structure consists of three alpha helices, by minimizing Z-score of the native state with respect to the misfolded ensemble, which is computed as $(E_{\text{nat}} - \langle G_{\text{misfold}} \rangle)/\sigma(G_{\text{misfold}})$ [47]. They were able to experimentally verify that one of the designed sequences folds into a native-like structure.

In the following, protein sequences are designed that are optimally stable in the folding energy model employed here. Of particular interest is the influence of contact correlations on the designed sequences. As discussed in the introduction, the interaction energy $U(a,b)$ of two amino acids a and b can be approximated by the product of the hydrophobicities of the two amino acids $-\epsilon_H\, h(a)h(b)$, that is, the problem of finding the energetically optimal sequence is mapped on the problem of finding the optimal hydrophobicity profile (HP). The solution was found to be the Effective Connectivity (EC), which maximizes the quadratic form eq. (1.9) under constraints on $\left[h\right]$ and $\left[h^2\right]$. The two constraints on the EC can be interpreted in two different ways. First, the constraints can be attributed to the mutation process which produces a random sequence with an average hydrophobicity value, which is equivalent to constraining the mean profile. The constraint on the mean squared profile values $\left[x^2\right]$ is equivalent to constrain the standard deviation of the HP, modeling the entropic force of the mutation process, i.e., the mutational drift towards sequences with a diverse amino acid content. Second, the mean hydrophobicity is highly correlated to the mean interaction energy $\left[U\right]$ (see Fig. 1.5, $cc = 0.77$) and the mean squared hydrophobicity $\left[h^2\right]$ is strongly correlated to $\left[U^2\right]$ ($cc = 0.94$, data not shown). Consequently, the constraints on the profile are equivalent to constraining $\left[U\right]$ and $\left[U^2\right]$, i.e., constraints on the misfolded ensemble are formulated using the REM. However, the two interpretations are not equivalent, as the first refers to the mutational process while the second refers to the selection against misfolded structures.

In the last Section it was shown that contact correlations have a significant contribution to the stability of sequences against misfolding. Hence, it is of interest to investigate whether one can exploit that knowledge to construct more stable sequences. For this purpose, the free energy difference ΔG is formulated in the hydrophobic approximation and the optimal HP is determined under the constraints that are applied on the EC. The scale of the hydrophobicity in the quadratic form (1.9) is not relevant. In the free energy difference the second cumulant introduces terms that have a product of two interaction energies, and hence a term that is to the power of four in hydrophobicity, making the scale of the HP relevant.

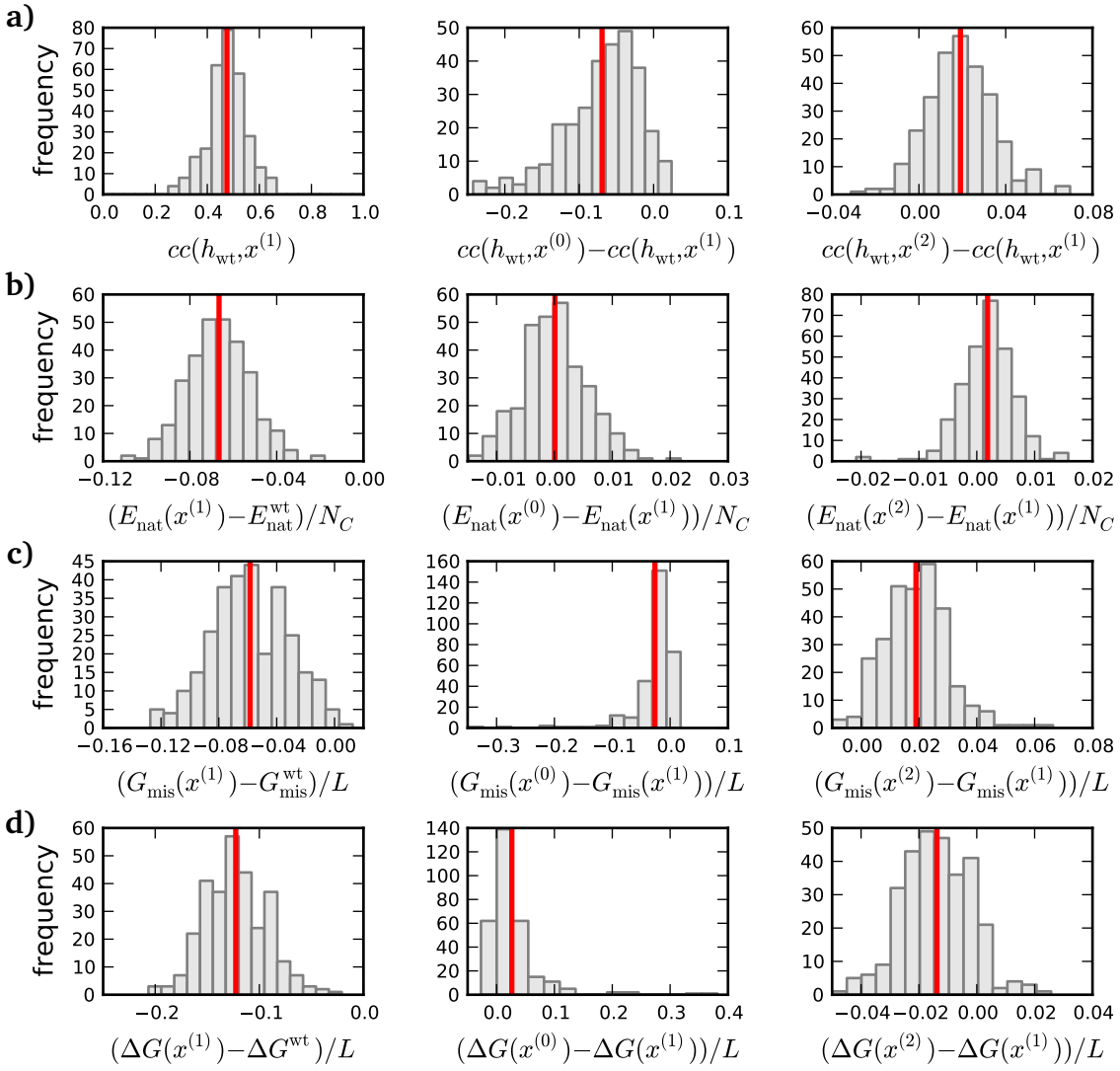

Figure 2.23: Assessment of the optimal hydrophobicity profiles PE, EC and CF-EC. The second and third rows show the difference of the PE and CF-EC with respect to the EC, which serves as a reference. The figures show histograms of **a)** the correlation coefficient of the profile with the HP of wild type sequences, **b)** free energy of the native state , **c)** free energy of the misfolded ensemble at $T = 1.2$, and **d)** free energy difference $\Delta G = E_{\rm nat} - G_{\rm misfold}$ at $T = 1.2$ computed for the sequences derived from the optimal profiles. The red line indicates the position of the mean.

To resolve this, the hydrophobicity h_i is substituted by the product of a profile x_i and the mean hydrophobicity $\left[h\right]$. Now the constraint $\left[x\right] = 1$ implies that the HP is correctly normalized to its mean value. Instead of making the mean hydrophobicity a freely adjustable parameter that has to be optimized together with the profile, it is estimated from the mean hydrophobicity of the wild type sequence of the protein.

As argued in the introduction, the interaction energy is best replaced by the hydrophobicities plus a repulsive term $-\epsilon_H h(a)h(b) + U_{\rm rep}$. The entropy of the misfolded ensemble does not

depend on the sequence and can therefore be neglected. Therefore, the free energy difference in terms of the profile **x** is written as

$$\frac{2\Delta G}{\epsilon_H \left[h\right]^2} \approx \sum_{ij} \left(C_{ij}^{\text{nat}} - \left\langle C_{ij} \right\rangle \right) x_i x_j + \frac{\epsilon [h]^2}{4 k_{\text{B}} T_{\text{prof}}} \sum_{ijkl} S_{ijkl} x_i x_j x_k x_l + \frac{U_{\text{rep}}}{2 k_{\text{B}} T_{\text{prof}}} \sum_{ij} D_{ij} x_i x_j + \text{const.} \quad , \tag{2.27}$$

where the temperature T_{prof} is introduced, which can be seen as a parameter and not necessarily identical to the physical temperature. The sum of each index is extended over all residues, which gives rise to an additional prefactor of $1/2$ to each sum. The U_{rep} term in the hydrophobic approximation gives rise to the D_{ij} in the free energy. To assess the contribution of different levels of contact correlations, a number of profiles $x_i^{(n)}$, which take into account more and more properties of the misfolded free energy, are defined.

In principle, the contribution of second cumulant in eq. (2.27) should be sufficient to produce normalized profiles. However, tests showed that profiles, which result from a maximization of eq. (2.27) without constraints, exhibit unwanted properties. Therefore, with the exception of the profile $x^{(0)}$, all profiles are derived with the constraint of the standard EC, i.e., $\left[x\right] = 1$ and $\left[x^2\right] = B$:

1. The profile $x_i^{(0)}$ is PE of the contact matrix. It minimizes only the native energy, without constraining $\left[x\right]$, fully neglecting negative design, i.e., it maximizes the quadratic form $\sum_{ij} C_{ij}^{\text{nat}} x_i x_j$, with the constraint $\sum_i x_i^2 = 1$.

2. The profile $x_i^{(1)}$ is the standard EC, which maximizes the same quadratic form $\sum_{ij} C_{ij}^{\text{nat}} x_i x_j$, but with both constraints $\left[x\right] = 1$ and $\left[x^2\right] = B$.

3. The profile $x_i^{(2)}$ maximizes the first term in eq. (2.27), $\sum_{ij} \left(C_{ij}^{\text{nat}} - \left\langle C_{ij} \right\rangle \right) x_i x_j$, i.e., it takes into account the contact frequency. It is referred to as the contact-frequency-corrected EC, CF-EC. $x_i^{(1)}$ and $x_i^{(2)}$ only depend on the native structure and not on the sequence, since they do not depend on the mean hydrophobicity parameter $\left[h\right]$.

4. The profile $x_i^{(c2)}$ takes into account two site correlations $S_{ijij} = \left\langle C_{ij} \right\rangle - \left\langle C_{ij}^2 \right\rangle$. It is referred to as the contact-correlation-pairs-2-corrected EC, CCP2-EC. The quantity to be maximized is quartic in x, which is approximated by a quadratic form, where the profile is approximated by the solution $x_i^{(2)}$. Thus, one maximizes $\sum_{ij} \left(C_{ij}^{\text{nat}} - \left\langle C_{ij} \right\rangle + \frac{\epsilon [h]^2}{4 k_{\text{B}} T_{\text{prof}}} S_{ijij} x_i^{(2)} x_j^{(2)} \right) x_i x_j$.

5. In the following step, three-sites correlations S_{ijik} are considered. The profile $x_i^{(c3)}$ is referred to as the contact-correlation-pairs-3-corrected EC, CCP3-EC and maximizes the quadratic form $\sum_{ij} \left[C_{ij}^{\text{nat}} - \left\langle C_{ij} \right\rangle + \frac{\epsilon [h]^2}{4 k_{\text{B}} T_{\text{prof}}} \left(S_{ijij} x_i^{(2)} x_j^{(2)} + \sum_{k \neq j} S_{ijik} x_i^{(2)} x_k^{(2)} + \sum_{k \neq i} S_{ijkj} x_k^{(2)} x_j^{(2)} \right) \right] x_i x_j$.

6. The profile $x_i^{(c4)}$ considers the whole correlation matrix S_{ijkl} and is referred to as the contact-correlation-corrected EC (CC-EC). It is found by maximizing the quadratic form $\sum_{ij}\left(C_{ij}^{\mathrm{nat}} - \left\langle C_{ij}\right\rangle + \frac{\epsilon[h]^2}{4k_{\mathrm{B}}T_{\mathrm{prof}}}\sum_{kl} S_{ijkl}x_k^{(2)}x_l^{(2)}\right)x_i x_j$.

7. Finally, the term proportional to U_{rep} is considered. The profile $x_i^{(c5)}$ maximizes

$$\sum_{ij}\left(C_{ij}^{\mathrm{nat}} - \left\langle C_{ij}\right\rangle + \frac{\epsilon[h]^2}{4k_{\mathrm{B}}T_{\mathrm{prof}}}\sum_{kl} S_{ijkl}x_k^{(2)}x_l^{(2)} + \frac{U_{\mathrm{rep}}}{2k_{\mathrm{B}}T_{\mathrm{prof}}}D_{ij}\right)x_i x_j$$

and is referred to as the contact-correlation-corrected-U_{rep} EC (CCU-EC).

Profiles, which take the second cumulant into account, are the solution of a quartic optimization problem, for whose solution no practical algorithm is known. A simple iterative scheme was tried but did not converge, so it was decided to compute the profile from the quadratic form, which approximates the quartic form, as discussed above.

The hydrophobicity profiles are assessed in two ways: First, it is assumed that wild type sequences have an nearly optimized stability. Thus, the optimal hydrophobicity is expected to be well correlated with the hydrophobicity profile $h(A_i)$ of the wild type sequence. Second, sequences are constructed from the optimal hydrophobicity and assess the change in free energy difference ΔG.

There is no clear strategy how to construct a sequence from a structural profile. It is assumed that the wild type sequence has an amino acid content that is optimal for the stability of the fold. Then, the optimal sequence is found from reshuffling the amino acids of the wild type sequence, such that the resulting hydrophobicity profile is optimally correlated with the structural profile. That is done by ranking sites by their values of the structural profile and assign an amino acid of the wild type sequence with the highest hydrophobicity to the site with the highest profile value. Then, the amino acid with the second highest hydrophobicity is assigned to the site with the second highest profile value and so on, until the least hydrophobic amino acid is assigned to the site with the lowest profile value.

Sequences, that are constructed in this way, differ largely from wild type sequences. Only about 11% of the amino acids in the reshuffled and wild type sequence are in the same position in the sequence, which is only marginally more than 7%, found for randomly reshuffled sequences. However, the profiles bear a significantly large similarity to the wild type HP. The correlation coefficient of the EC and the HP of wild type sequences is on average $\langle cc(h_{\mathrm{wt}}, x^{(1)})\rangle = 0.476 \pm 0.004$ (see Fig. 2.23.a)). Concentrating first on the profiles that do not depend on temperature, it is found that the PE has significantly lower correlation ($\langle cc(h_{\mathrm{wt}}, x^{(0)})\rangle = 0.406 \pm 0.005$), as concluded in previous studies. With respect to the EC, the CF-EC can indeed increase the correlation by a small but significant amount to $\langle cc(h_{\mathrm{wt}}, x^{(2)})\rangle = 0.495 \pm 0.004$.

More interesting than the correlation with the wild type sequences is the stability of the constructed sequences (see Fig. 2.23.b)-d)). In relation to wild type sequences, the EC decreases the free energy of the native state for all proteins in the test set, but also decreases the misfolded free energy. The improvement of the native free energy outweighs the decrease in misfolded energy and the EC-sequences have an improved stability ΔG.

The PE and CF-EC-sequences yield a native free energy that is on average the same as for the EC-sequences. However, the misfolded free energies differ largely. The free energy difference

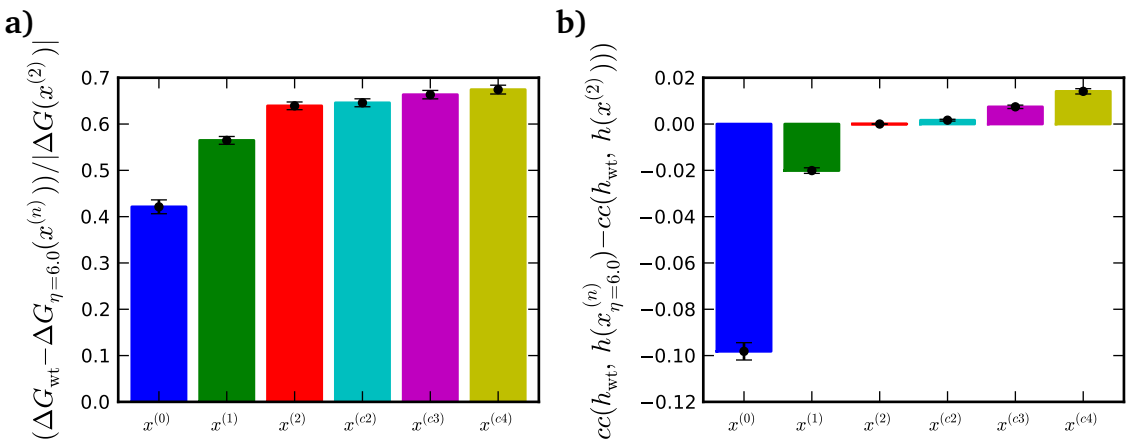

Figure 2.24: Free energy difference of wild type and optimal sequences (**a**) and correlation of optimal hydrophobicity profiles with hydrophobicity of wild type sequences (**b**).

ΔG with respect to the EC is strongly impaired for the PE, mostly due to some outliers. In contrast, it improves for the CF-EC for most of the sequences. Thus, the consideration of the contact frequency in the misfolded ensemble improves ΔG by increasing G_{misfold}, that is improving the stability against misfolding, while keeping the stability against unfolding, i.e. E_{nat}, almost constant.

Profiles that account for contact correlations depend on the temperature T_{prof}, giving rise to a whole series of profiles. The temperature can be seen as a parameter that controls the influence of contact correlations, which is strongly increased with decreasing temperature. However, the temperature has also a physical meaning, which has to be respected. In fact, if the temperature T_{prof} is very low, it can be below the freezing temperature of the resulting sequence. However, in this case the free energy is evaluated at freezing temperature and consequently the free energy that is used in the profile computation is wrong. This should yield unstable sequences, which, however, could not be observed, as discussed below. In addition, a freezing temperature cannot be easily defined for a profile, as cumulants are only computed in the hydrophobic approximation. Therefore, the temperature T_{prof} is viewed as parameter.

The inverse temperature is varied from one energy unit, which is approximately room temperature, to 12 energy units. At infinite temperature, the temperature dependent profiles coincide with the CF-EC. If the temperature T_{prof} is decreased, the correlation coefficient of the profiles with the HP of wild type sequences increases slightly. The order of the improvement increases with the amount of contact correlation that are taken into account (see Fig. 2.25.a)). That is, the increase is larger for the CC-EC than for the CCP2-EC or CCP3-EC, supporting the idea that the consideration of contact correlations is important in negative design. The correlation coefficient for the CC-EC has a maximum at $T_{\mathrm{prof}} = 1/7$. If the U_{rep} term is considered, the correlation with the wild type is slightly decreased.

Again, sequences are constructed from the profiles and compute the free energy at the room temperature $T = 1.2$ and at the temperature of the profile construction T_{prof}. In the latter case, below a rather high temperature of $T = 0.5$ the profile temperature T_{prof} is below the freezing temperature of the constructed sequences, so the free energy is effectively computed at freezing temperature below $T = 0.5$. The curves for profile and room temperature are hardly different

(compare Fig. 2.25.b) and c)), because room temperature is only marginally larger than freezing temperature of the chains. Therefore, they can be discussed in parallel.

The free energy is compared to the CF-EC, which produced the most stable sequences so far, as discussed above. With decreasing profile temperature T_{prof} the influence of negative design by contact correlations is increased and the free energy of the native state as well as for the misfolded ensemble is increased as expected. Again, the effect is stronger the more terms of the second cumulant are considered. The U_{rep} term, however, attenuates the effect. The net effect is an improvement of the free energy difference ΔG with respect to the CF-EC. Indeed, the increase is larger the more terms of contact correlations are considered. However, the effect is rather small. The CC-EC reaches its maximal improvement at $T = 1/6$ with only 2.5% at room temperature and 3.5% at freezing temperature. The consideration of the repulsive term U_{rep} diminishes the improvement, in accordance with the small correlation coefficient of the profile $x^{(5)}$ with the wild type HP.

The relatively low optimal temperature, which is much below freezing temperature, suggests that the scheme of profile construction is not yet optimal. The current approach can be seen as the first step in an iterative scheme and the current solution is going into the right direction, such that the solution can be increased if the step size is increased, i.e., if a smaller temperature is applied.

The results are summarized in Fig. 2.24 for the temperature $T_{\text{prof}} = T = 1/6$. All profiles produce sequences that are more stable than wild type sequences. With respect to the EC, the new profiles can improve the stability. The consideration of contact frequency brings about the largest effect while contact correlations yield only a small but significant improvement. This finding is reflected in the correlation coefficient with the HP. Indeed, the profiles improve in two, apparently imposing ways: They become closer to wild type sequences in terms of the correlation coefficient with the wild type HP and, at the same time, become seemingly dissimilar from wild type sequences as their stability improves with respect to the stability of wild type sequences. The largest improvement results from the consideration of contact frequency. The corresponding change in the profile is specific in negative design and is described in the next Section.

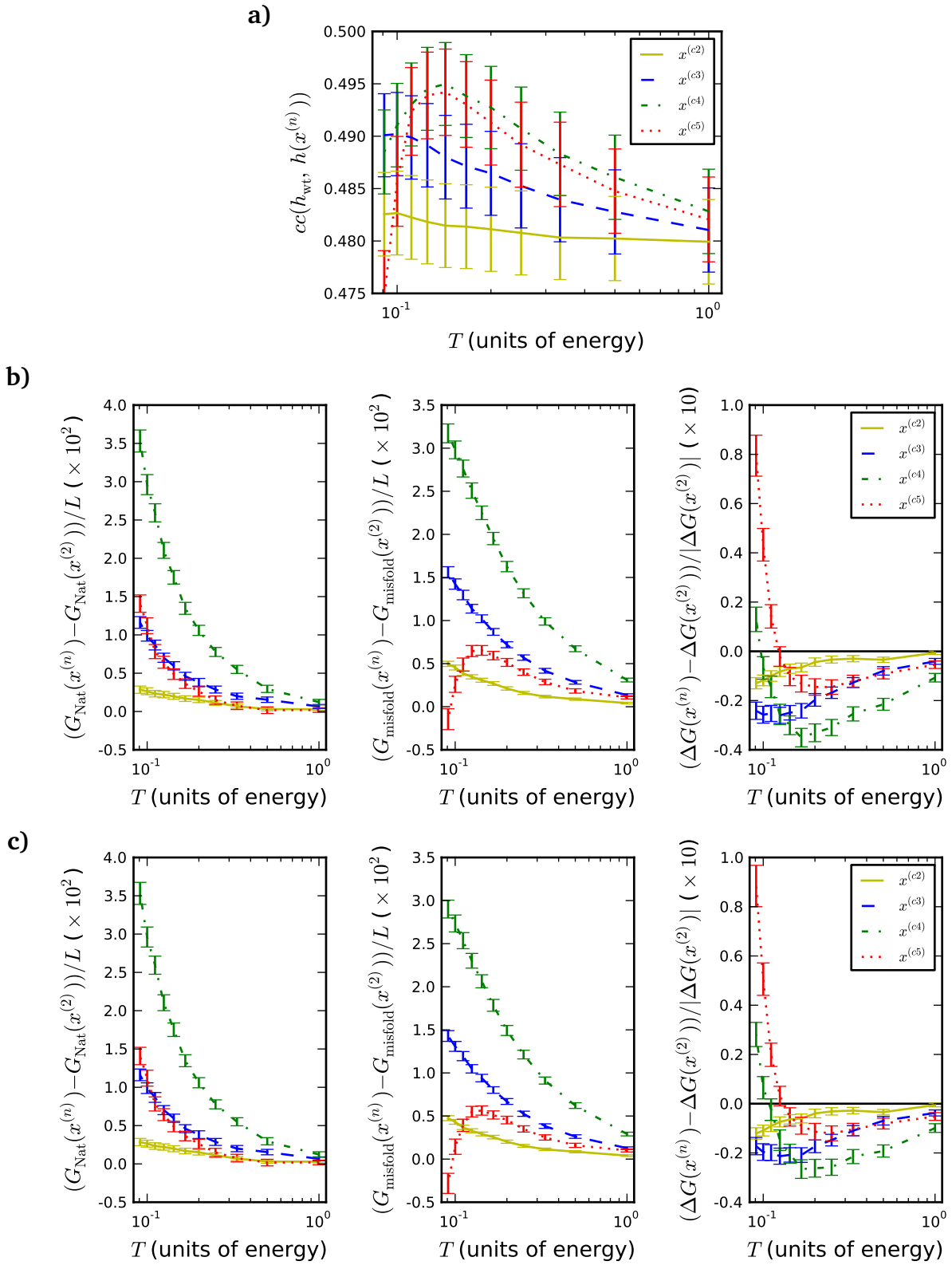

Figure 2.25: Correlation with wild type hydrophobicity **(a)** and free energy difference versus temperature of profile sequences. **b)** The temperature for the computation of the profile and for the free energy of the sequences are equal. **c)** The temperature for the computation of the free energy is held fixed at $T = 1.2$.

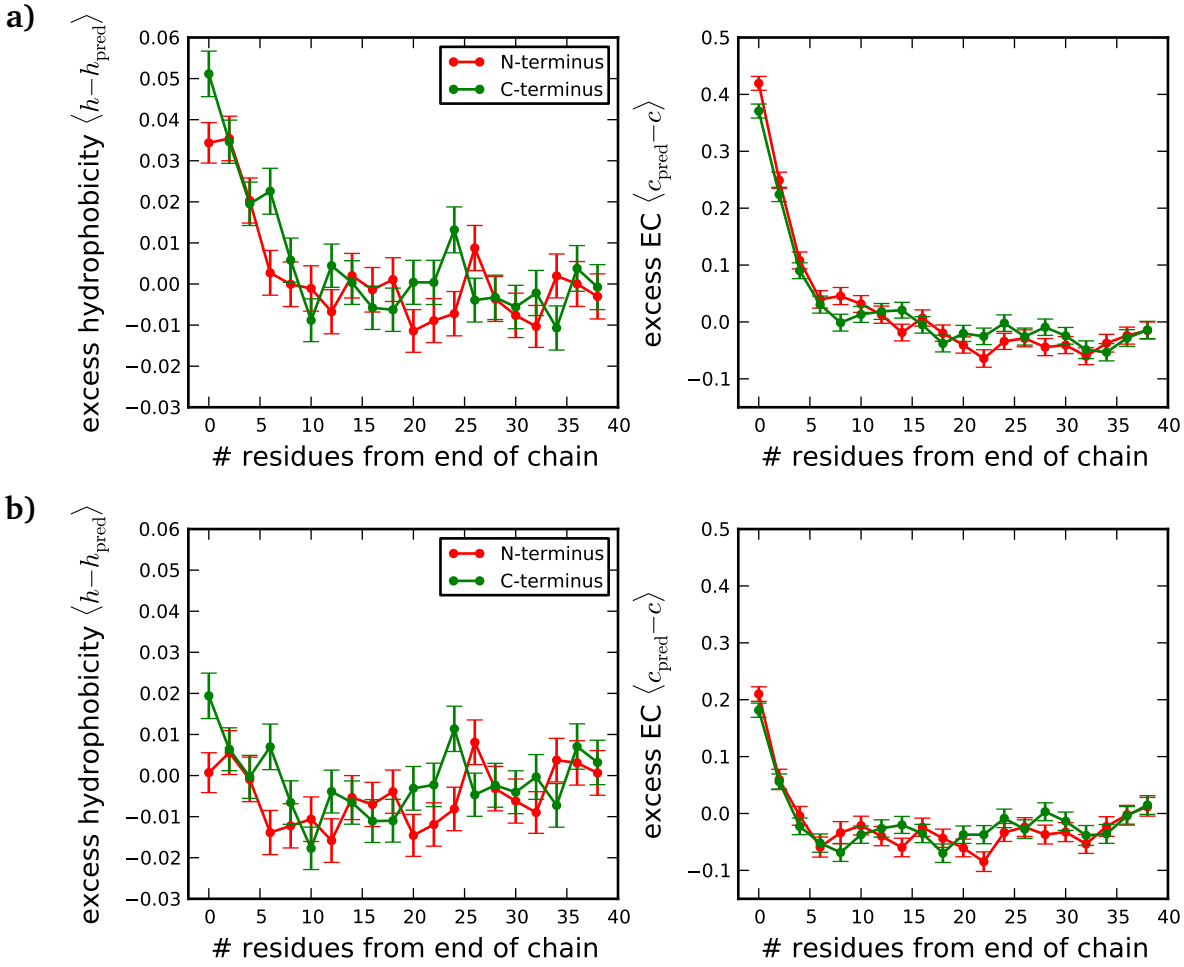

Figure 2.26: Excess of hydrophobicity and profile value for EC (**a**) and CF-EC (**b**). The excess hydrophobicity is computed as the difference of the hydrophobicity predicted from a fit to the profile values and the obeserved hydrophobicity (left column) and vice versa (right column). **a)** At the termini hydrophobicities are more hydrophobic than expected from the EC. **b)** This effect is reduced by the consideration of contact frequency in the CF-EC.

Chain end effect

The largest change of the EC results from the consideration of contact frequency in the CF-EC. A closer inspection reveals, that the change consists predominately in an increase of the EC at the termini of the chains. Indeed, most of the increase of the correlation with the hydrophobicity is caused by this change. The hydrophobicity of the amino acids is underestimated by the EC, which can be easily seen by a simple test. The hydrophobicity i is fitted to the EC c by the linear fit,

$$h_i^{\mathrm{pred}} = a\, x_i + b \tag{2.28}$$

where a and b are fit parameters. From the fit one can predict the hydrophobicity and compute the excess hydrophobicity, which is the difference of the prediction h_i^{pred} and the observed hydrophobicity $h(A_i)$. Fig. 2.26 shows the average excess hydrophobicity binned by the distance

to the termini. Clearly, the EC underestimates hydrophobicity at the end of the chains. If the contact frequency is considered, the excess hydrophobicity is reduced.

This finding can be interpreted in the sense of negative design. The average number of contacts of the residues i in the misfolded ensemble is $\sum_j \left\langle C_{ij} \right\rangle$. The contact frequency is constant throughout the inner part of the chain, and decreases towards the termini, because terminal residues lack possible contact partners beyond the end of the chain. Thus, terminal residues contribute less to the misfolded free energy and positive design can increase their hydrophobicity to strengthen their native contacts.

This is reflected by the different behavior of the EC and the CF-EC. The EC does not know about the properties of the misfolded ensemble, while the CF-EC maximizes the difference between native energy and mean energy of misfolded ensemble $C_{ij}^{\mathrm{nat}} - \left\langle C_{ij} \right\rangle$.

2.5 Discussion

The most widely used theoretical description of the misfolded ensemble is the REM. Although the REM provides a good estimate of the misfolding free energy, it neglects the contact frequency and contact correlations, which have important contributions. In combination with a cumulant expansion, the characterization of contact frequency and correlation yields a good approximation of the misfolded free energy measured from threading. The improved description of the misfolded free energy enabled the detection of negative design in wild type sequences. The null model was represented by randomized sequences, i.e., sequences that are randomly drawn from a background distribution and shuffled wild type sequences.

The comparison to random sequences turned out to be problematic as random sequences cover a broad range from very hydrophobic to very hydrophilic sequences, whereas shuffled sequences provide a more meaningful comparison. Their free energy is narrowly distributed about the free energy of the corresponding wild type sequences, agreeing with the REM, which predicts the same misfolded free energy for all shuffled sequences. The small difference between the shuffled and wild type sequences is, however, significant and in favor for negative design in wild type sequences.

The analyses are attenuated by the fact that wild type sequences have to be optimized for positive as well as for negative design. The signals for negative design become much clearer when the negative design scores were evaluated for native and non-native contacts separately. Such a splitting is unfortunately not possible for the free energy, as for the misfolded free energy contacts and non-contacts cannot be disentangled. Thus, a more direct evidence for negative design is not possible.

The largest contribution to the signal comes from the first and second cumulant, where the second cumulant shows the strongest difference with respect to the shuffled ensemble, suggesting that wild type sequences exploit contact correlations to improve their negative design. However, wild type sequences have also be optimized with respect to positive design as well. This can be seen as an indication that positive and negative design are to a certain extent not frustrated. Furthermore, this may also explain why the consideration of contact correlations did only improve the designed sequences by a minor amount. However, that is only an observation at the moment and needs further tests.

The analysis considered the average effect of all residues on negative design. What is also interesting is to investigate a signal for individual pairs. However, this can only be achieved by an evolutionary average over many sequences, similar to the analysis of Horovitz *et al.*, since a single sequence is not necessarily optimally stable.

The energy model used for the detection of negative design is fitted such that wild type sequences are maximally stable against misfolding, which might arouse the suspicion that the signal of negative design is due to some over-fitting of the model. As a simple test, the free energy and the negative design scores where computed with a different $U(a,b)$, which is computed from a Boltzmann inversion of the contacts frequency of $U'(a,b) = -\ln \frac{P(a,b|c=1)}{P(a,b|c\neq -1)}$ (cf. eq. (3.46) on page 73). This interaction matrix U' is highly correlated with the standard matrix, while it is still significantly different. The tests of the free energy and the negative design scores produced very similar results. Thus, an over-fitting can be ruled out.

For sequence design new variants of the EC are defined, which take into account negative design more explicitly. The largest effect has the inclusion of the contact frequency in the

definition of the EC. Even though it effects only the terminal residues, it improved the correlation with HP of wild type sequences significantly. The effect of contact correlations is very small and can only be detected by averaging many sequences. Indeed, the temperature parameter used in the profile construction has to be lowered to values that are far below the physical freezing temperature in order to optimize the contribution of the second cumulant.

The reasons are related to the manifold approximations in the profile and sequence construction: First, the EC provides only an estimate of the free energy in the hydrophobic approximation. Second, the solution of the maximization problem is only approximately solved. Third, the constraints of the EC might not be optimal. However, loosening the constraints and leave the normalization to the second cumulant did not improve the results.

3 Correlated Mutations

3.1 Introduction

The folding and function of a protein is a cooperative phenomenon that arises from the physical interaction of amino acids in native and non-native states. Thus, it is natural to assume that the probability of the acceptance of an amino acid mutation at one site in protein depends on the other sites, i.e., amino acid substitutions are correlated. Even though the observed correlation refers to the amino acid *substitutions*, this field of study is commonly referred to as correlated *mutations*.

The correlation of mutations can be understood from the probability of a sequence to occur during evolution. In particular, the correlation is defined as the relation of the pair-specific probability $P_{ij}(a,b)$ of the amino acids a and b to occur at sites i and j, respectively, to its random expectation, which is given by the product of the site-specific probabilities $P_i(a)P_j(b)$. In this thesis the measure for the correlation is the ratio of these probabilities,

$$Q_{ij}(a,b) = \frac{P_{ij}(a,b)}{P_i(a)\,P_j(b)} \quad . \tag{3.1}$$

If two amino acids are uncorrelated, the pair probability $P_{ij}(a,b)$ is equal to the product of the site-specific frequencies $P_i(a)\,P_j(b)$ and Q becomes one. If two amino acids occur more often than the random expectation, i.e., the amino acids are positively correlated, Q is greater than one.

A widely used measure for the overall correlation of two sides i and j is the mutual information (MI), which is defined as

$$\mathrm{MI}_{ij} = \sum_{a,b} P_{ij}(a,b) \log \frac{P_{ij}(a,b)}{P_i(a)\,P_j(b)} = \sum_{a,b} P_{ij}(a,b) \log Q_{ij}(a,b) \quad . \tag{3.2}$$

The MI measures the average information that is acquired about site i, when one observes which amino acid is found at site j and vice versa. If the two sites are uncorrelated, no information can be acquired about the other site and the MI is zero. The mutual information can be written as the difference of the pair-specific entropy minus the entropy of the site-specific probabilities,

$$\begin{aligned} \mathrm{MI}_{ij} &= -\sum_{a,b} P_{ij}(a,b) \ln P_{ij}(a,b) + \sum_{a} P_i(a) \ln P_i(a) + \sum_{b} P_j(b) \ln P_j(b) \\ &= S(P_{ij}(a,b)) - S(P_i(a)) - S(P_j(b)) \end{aligned} \tag{3.3}$$

Correlated mutations of amino acids can be measured from evolutionary data of a protein, contained in the alignment of many different sequences, which arose during evolution. The thermodynamic stability imposes strong constraints on sequence evolution. In particular, the interaction of two residues that form a native contacts imposes particularly strong constraints on the coevolution of the residues, which was exploited to predict native contacts between residues

from the observation of correlated mutations from evolutionary data [48, 49, 50, 51, 52]. Such a prediction is not restricted to single proteins, as the function of proteins often relies on the interaction with other proteins. The analysis of correlated mutations of two different proteins that are known to interact have successfully been used to predict the residues that partake in the binding of the two proteins, i.e., that are close in space in the functional arrangement of the two proteins [53].

However, the observation of strong correlations between residues, which are distant in the native protein structure, led to the idea that also the interaction of residues in the misfolded ensembles can give rise to correlated substitutions [31, 54].

As another cause for correlations functional mechanism were identified. Clusters of correlated sites, which extend over a large part of the protein structure, were found and could be attributed to allosteric communication in proteins [55, 56, 57]. In allosteric communication the information of binding of a ligand at one site of the protein is transduced to a distant other site, where it causes structural deformations that modify the binding specificity of the functional site of the protein. This mechanism needs the cooperative interactions of amino acids throughout the structure, which are reflected in the correlated mutations of the respective sites. In a different study correlated mutation analysis was used to identify functionally important residues, which were found to be hub nodes in a network of correlated sites [58].

All these studies give important information about correlated mutations, but they only measure and interpret empirical correlations. A model that quantitatively describes correlations is still missing. Unlike protein specific functional constraints, the thermodynamic stability of the native state is a generic selection criterion for most proteins. Thus, the stability as a selection criterion for protein sequence evolution allows to formulate such a model, which will be investigated in this chapter. The model exploits a formal analogy between statistical physics and evolution, which has been found by Sella and Hirsh [59]: The entropy is represented by the sequence entropy, which mutations attempt to maximize, while the energy is represented by the fitness of the phenotype, which is controlled by natural selection. Thus, the model maximizes a free energy, consisting of sequence entropy and the stability of the fold, which is connected to the fitness. The solution of minimal free energy yields a prediction for the correlation $Q_{ij}(a, b)$. The model is tested for sequences generated by a simulation of protein evolution and by an statistical analysis of wild type sequences and structures.

3.2 Theory

The project was done in close collaboration with Ugo Bastolla, who developed the theory of correlated mutations in the context of this collaboration. Since the theory has not been published yet and is crucial to the understanding, all analytical computations are summarized in this thesis.

3.2.1 Maximal entropy approach

During evolution a protein can adopt an enormously large number of sequences by random mutations. Therefore, it is advisable to approach protein sequence evolution from a statistical point of view by considering the probability $P_{1,\ldots,L}(A_1, \ldots, A_N)$ that a amino acid sequence $A_1, \ldots, A_N$ occurs during evolution. Here, it is assumed that mutations take place on the DNA level, where

they occur randomly and independently for all nucleotides. Such a random process has an equilibrium distribution of nucleotides and hence codons, which translates to a distributions of amino acids by means of the genetic code. Then, an amino acid a will occur in the sequence with the background probability $P_{\mathrm{bg}}(a)$ and the background probability for the entire amino acid sequence is then the product of all single site probabilities, that is $\prod_{i=0}^{L-1} P_{\mathrm{bg}}(A_i)$. In other words, mutations tend to maximize the relative entropy of sequence distribution $S(P_{1,\ldots,L}(A_1, \ldots, A_N))$, which is defined as

$$S(P_{1,\ldots,L}(A_1, \ldots, A_N)) = -\sum_{A_i,\ldots,A_N} P_{1,\ldots,L}(A_1, \ldots, A_N) \ln \frac{P_{1,\ldots,L}(A_1, \ldots, A_N)}{\prod_{i=0}^{L-1} P_{\mathrm{bg}}(A_i)} \tag{3.4}$$

Apparently, maximum entropy is attained if the sequence distribution is equal to the background distribution.

However, selection imposes constraints on sequence space, and the assumption is that sequence evolution is constrained by protein folding thermodynamics. These constraints can be formulated in a couple of different ways. For example, the free energy difference between the folded and misfolded ensemble could be constrained by the free energy difference between the native state and the misfolded ensemble as discussed in the previous Chapter. However, this would again confront us with the problem of a properly defined temperature. Instead, two separate constraints on unfolding and misfolding are formulated. The former is formulated by constraining the free energy of the native state E_{nat}, and the latter by constraining the minimal free energy of the misfolded ensemble E_{min}. In particular, the restriction could consist in thresholds for the two energies or a fitness function that depends on two energies and should acquire a certain mean. The former is difficult to implement and for the latter it is not clear which fitness function one should take.

What is more general is to constrain the evolutionary average of E_{nat} and E_{min} separately. An evolutionary average is taken with the probability distribution $P_{1,\ldots,L}(A_1, \ldots, A_N)$ of the sequences that occur during evolution. Since the free energy of the native state is computed in the model by a sum over all pairs of residues, correlations of only two sites contribute to the evolutionary average $\langle E_{\mathrm{nat}} \rangle_{\mathrm{evol}}$

$$\begin{aligned}\langle E_{\mathrm{nat}} \rangle_{\mathrm{evol}} &= \sum_{A_1,\ldots,A_N} \sum_{i<j-2} C_{ij}^{\mathrm{nat}} U(A_i, A_j) P_{1,\ldots,L}(A_1, \ldots, A_N) \\ &= \sum_{i<j-2} \sum_{a,b} C_{ij}^{\mathrm{nat}} U(a, b) P_{ij}(a, b) = \sum_{i<j-2} \sum_{a,b} C_{ij}^{\mathrm{nat}} U(a, b) P_i(a) P_j(b) Q_{ij}(a, b)\end{aligned} \tag{3.5}$$

A formulation amenable to analytical treatment of the minimal energy of the misfolded ensemble is given by the estimate of the Gaussian extreme value statistics (as discussed in Section 2.3 on page 17)

$$E_{\mathrm{min}}^{(\mathrm{REM})} = \langle E \rangle - \sigma(E) \sqrt{2 \ln N_{\mathrm{mis}}} \tag{3.6}$$

$$\text{with} \quad \langle E \rangle = \sum_{i<j-2} \left\langle C_{ij} \right\rangle U(A_i, A_j) \text{ and } \quad \sigma(E) = \sqrt{\sum_{i<j-2} \sum_{k<l-2} S_{ijkl} U(A_i, A_j) U(A_k, A_l)} \quad . \tag{3.7}$$

Apparently, the evolutionary average of $E_{\min}$ is the difference of the averages of $\langle E \rangle$ and $\sigma(E)$. To assure analytical tractability the average $\sigma(E)$ is computed as the square root of the average $\sigma^2(E)$, otherwise the average would require the correlation of all sites

$$\langle\langle E\rangle\rangle_{\text{evol}} = \sum_{i<j-2}\sum_{a,b}\left\langle C_{ij}\right\rangle U(a,\,b)P_{ij}(a,b) = \sum_{i<j-2}\sum_{a,b}\left\langle C_{ij}\right\rangle U(a,\,b)P_i(a)\,P_j(b)\,Q_{ij}(a,b) \tag{3.8}$$

$$\sqrt{\left\langle\sigma^2(E)\right\rangle_{\text{evol}}} = \sqrt{\sum_{i<j-2,k<l-2}\sum_{a,b,c,d} S_{ijkl}U(a,\,b)\,U(c,d)\,P'_{ijkl}(a,b,c,d)} \quad , \tag{3.9}$$

where $P'_{ijkl}(a,b,c,d)$ refers to the probability $P_{ijkl}(a,b,c,d)$ if the four indices i, j, k, and l are different, and to the probability $P_{ijk}(a,b,c)$ if two indices are identical and so on.

Finally, it is necessary to ensure that the probability distribution is normalized to one, that is, the relation $\sum_{A_1,\ldots,A_N} P_{1,\ldots,L}(A_1,\ldots,A_N) = 1$ has to be fulfilled. In order to find the probability distribution that maximizes the entropy subject to the constraints formulated so far, the method of Lagrange multipliers is applied. To this end, the Lagrange multiplier Λ_E is introduced, which regulates the constraint on $\left\langle E_{\text{nat}}\right\rangle_{\text{evol}}$, the multiplier Λ_e, which regulates the constraints on $\left\langle E_{\min}^{(\text{REM})}\right\rangle_{\text{evol}}$ and the multiplier Λ_0, which regulates the constraint on the normalization of the probability distribution. Thus, the following function has to be maximized with respect to $P_{1,\ldots,L}(A_1,\,\ldots,\,A_N)$,

$$\begin{aligned} F(P_{1,\ldots,L}(A_1,\,\ldots,\,A_N);\,\Lambda_E,\,\Lambda_e,\,\Lambda_0) =& S(P_{1,\ldots,L}(A_1,\,\ldots,\,A_N)) - \Lambda_0\left(\sum_{A_1,\ldots,A_N} P_{1,\ldots,L}(A_1,\,\ldots,\,A_N) - 1\right) \\ & - \Lambda_E\left(\left\langle E_{\text{nat}}\right\rangle_{\text{evol}} - \left\langle E_{\text{nat}}\right\rangle^0_{\text{evol}}\right) - \Lambda_e\left(\left\langle E_{\min}^{(\text{REM})}\right\rangle_{\text{evol}} - \left\langle E_{\min}^{(\text{REM})}\right\rangle^0_{\text{evol}}\right) \quad , \end{aligned} \tag{3.10}$$

where the Lagrange multipliers have to be adjusted to values such that the evolutionary averages of E_{nat} and $E_{\min}$ adopt their desired means $\left\langle E_{\text{nat}}\right\rangle^0_{\text{evol}}$ and $\left\langle E_{\min}^{(\text{REM})}\right\rangle^0_{\text{evol}}$, respectively. By finding the null of the derivative with respect to the sequence probability of eq. (3.10), the extremum of F is found to be

$$\begin{aligned} P_{1,\ldots,L}(A_1,\ldots,A_N) = \prod_{i=0}^{L-1} P_{\text{bg}}(A_i)\exp\Bigg(& -\Lambda_0 - \Lambda_E\sum_{i<j} U(A_i,A_j)C^{\text{nat}}_{ij} \\ & - \Lambda_e\Big(\sum_{i<j-2}\left\langle C_{ij}\right\rangle U(A_i,A_j) - \frac{1}{2}\left\langle\sigma(E)\right\rangle^{-1}_{\text{evol}}\sqrt{2\ln N_{\text{mis}}} \\ & \sum_{i<j-2,k<l-2} S_{ijkl}U(A_i,A_j)U(A_k,A_l)\Big)\Bigg) \quad . \end{aligned} \tag{3.11}$$

Due to the large number of possible sequences, maximizing the entropy over all sequences, which accounts for the full correlation contained in the probability distribution $P_{1,\ldots,L}(A_1,\ldots,A_N)$, is computationally very demanding. Therefore, a cluster expansion of the entropy is adopted, which considers the correlation of only subsets of all sites, the so called clusters [60, 61] (for a clear explanation of the cluster expansion in an information theoretical context see [62]).

3.2.2 Cluster expansion of free energy

In principle, any set of sites can form a cluster. The smallest possible cluster consists of one residue, which is equivalent to the assumption of independent single sites. The entropy in the cluster expansion of single sites is obtained by summing the relative entropy of the single site clusters,

$$S_1(\{P_i(a)\}) = -\sum_{i,a} P_i(a) \ln \frac{P_i(a)}{P_{\mathrm{bg}}(a)} \quad . \tag{3.12}$$

The mean energies are computed in the cluster expansion by substituting $Q_{ij}(a,b)$ in eq. (3.8) by one. In principle, one could replace the probability $P'_{ijkl}(a,b,c,d)$ in eq. (3.9) by the product of site-specific probabilities, resulting, however, in rather complicated derivatives. In the previous Chapter it was shown that the contribution of the S_{ijij}-term has the largest contribution to $\sigma(E)$. Thus, a sufficiently good approximation is obtained, if one restricts oneself to the S_{ijij}-term and replaces S_{ijkl} with $\sigma_{ij} \equiv \left\langle C_{ij} \right\rangle - \left\langle C_{ij} \right\rangle^2 \delta_{ik}\delta_{jl}$ (see Section 2.4.4 on page 34),

$$\begin{aligned} \sqrt{\left\langle \sigma^2(E) \right\rangle_{\mathrm{evol}}} &\approx \sqrt{\sum_{i<j-2} \sum_{a,b} \sigma_{ij} U^2(a,\, b)\, P_{ij}(a,b)} \\ &= \sqrt{\sum_{i<j-2} \sum_{a,b} \sigma_{ij} U^2(a,\, b)\, P_i(a)\, P_j(b)\, Q_{ij}(a,b)} \quad . \end{aligned} \tag{3.13}$$

Again, I obtain the approximation of independent sites if Q is set to one. Finally, to ensure the normalization of the site-specific probabilities, I introduce the Lagrange multipliers θ_i yielding the function to be minimized

$$\begin{aligned} F_1^{\mathrm{cluster}}(\{P_i(a)\}; \Lambda_E, \Lambda_e, \theta_i) = S_1(\{P_i(a)\}) &- \Lambda_E \left(\left\langle E_{\mathrm{nat}} \right\rangle_{\mathrm{evol}} (\{P_i(a)\}) - \left\langle E_{\mathrm{nat}} \right\rangle_{\mathrm{evol}}^0 \right) \\ &- \Lambda_e \left(\left\langle E_{\mathrm{min}}^{(\mathrm{REM})} \right\rangle_{\mathrm{evol}} (\{P_i(a)\}) - \left\langle E_{\mathrm{min}}^{(\mathrm{REM})} \right\rangle_{\mathrm{evol}}^0 \right) \\ &+ \sum_i \theta_i \left[\sum_a [P_i(a)] - 1 \right] \quad . \end{aligned} \tag{3.14}$$

From the null of the derivative of eq. (3.14) with respect to the single site probability $P_i(a)$, one finds an implicit equation for the maximum of F_1^{cluster}

$$\begin{aligned} P_i(a) = P_{\mathrm{bg}}(a) \exp \Bigg(&- \theta_i - \Lambda_E \sum_{k \neq i,b} C_{ij}^{\mathrm{nat}} U(a,b) P_j(b) \\ &- \Lambda_e \Big(\sum_{j \neq i,b} \left\langle C_{ij}^{\mathrm{nat}} \right\rangle - \frac{1}{2 \left\langle \sigma(E) \right\rangle_{\mathrm{evol}}} \sum_{j \neq i,b} \sigma_{ij} U(a,b) P_j(b) \Big) \Bigg) \quad . \end{aligned} \tag{3.15}$$

The solution of the implicit equation requires a numerical approach. Since a useful and much less complicated description of site-specific probabilities has already been formulated with structural profiles, as discussed in the introduction, the comparison of site-specific probabilities obtained from the cluster expansion and structural profiles is postponed to a later work.

To study correlations, clusters consisting of two sites are considered, thus neglecting correlations of three or more sites. Due to the flexibility of the chain every pair of sites can be interacting in a protein structure, both in the native state or a misfolded state, consequently being candidates for a high correlation. Therefore, the two-sites-clusters of all possible pairs of sites are considered. To begin with the cluster expansion of the entropy, the sum of the relative entropy of all pairs of sites is considered,

$$\begin{aligned} S_2'(\{P_{ij}(a,b)\}) &= -\sum_{i<j}\sum_{a,b} P_{ij}(a,b)\log\frac{P_{ij}(a,b)}{P_{\mathrm{bg}}(a)P_{\mathrm{bg}}(b)} \\ &= -\sum_{i<j}\sum_{a,b} P_{ij}(a,b)\log P_{ij}(a,b) + (L-1)\sum_{i,a} P_i(a)\log P_{\mathrm{bg}}(a) \end{aligned} \tag{3.16}$$

In a cluster expansion the contribution of clusters, which are intersections of larger clusters, has to be subtracted. The only intersection of two two-site clusters is a single site, if the two clusters have this site in common. It is easy to see that every site is contained $L-1$ times in all two-site-clusters. Hence, every site is over-counted $L-2$ times and the extra relative entropy has to be subtracted, yielding the correct relative entropy of the cluster expansion,

$$\begin{aligned} S_2(\{P_{ij}(a,b)\},\{P_i(a)\}) &\equiv -\sum_{i<j}\sum_{a,b} P_{ij}(a,b)\ln\frac{P_{ij}(a,b)}{P_{\mathrm{bg}}(a)P_{\mathrm{bg}}(b)} + (L-2)\sum_{i,a} P_i(a)\ln\frac{P_i(a)}{P_{\mathrm{bg}}(a)} \\ &= -\sum_{i<j}\sum_{a,b} P_{ij}(a,b)\ln P_{ij}(a,b) + (L-2)\sum_{i,a} P_i(a)\ln P_i(a) \\ &\quad + \sum_{i,a} P_i(a)\ln P_{\mathrm{bg}}(a) \quad . \end{aligned} \tag{3.17}$$

It is convenient to formulate the entropy in terms of the correlation measure Q, which is to be predicted. This is achieved by replacing the entropy of the two-sites clusters by the mutual information (eq. (3.2)). The mutual information can be written as the entropy of the two sites minus the entropy of the single sites. Thus, by summing the mutual information over all pairs, eq. (3.17) is obtained, but without the term of the the background probability $P_{\mathrm{bg}}(a)$ and the factor for the site-specific entropy is $L-1$ instead of $L-2$. Thus, by adding the sum over all sites of the site-specific relative entropy $-\sum_{i,a} P_i(a)\ln P_i(a)/P_{\mathrm{bg}}(a)$ to the summed mutual information, eq. (3.17) is obtained, which then reads

$$\begin{aligned} S_2(\{P_{ij}(a,b)\},\{P_i(a)\}) &= -\sum_{i<j}\sum_{a,b} P_{ij}(a,b)\ln\frac{P_{ij}(a,b)}{P_i(a)\,P_j(b)} - \sum_{i,a} P_i(a)\ln\frac{P_i(a)}{P_{\mathrm{bg}}(a)} \\ \Rightarrow S_2(\{Q_{ij}(a,b)\},\{P_i(a)\}) &= -\sum_{i<j}\sum_{a,b} P_i(a)\,P_j(b)\,Q_{ij}(a,b)\ln Q_{ij}(a,b) - \sum_{i,a} P_i(a)\ln\frac{P_i(a)}{P_{\mathrm{bg}}(a)} \quad . \end{aligned} \tag{3.18}$$

The evolutionary average of free energy of the native state (eq. (3.8)) relies already on two-site correlations and for $\langle\sigma(E)\rangle_{\mathrm{evol}}$ the approximation eq. (3.13) is adopted. With the cluster expansion performed, the variables of the function to be minimized change from the probability of the entire sequence $P_{1,\ldots,L}(A_1,\ldots,A_N)$ to the site-specific $P_i(a)$ and pair frequencies

$P_{ij}(a,b)$. Thus, the constraint on the marginalization of $P_{1,\dots,L}(A_1,\dots,A_N)$ has to be replaced by constraints that ensure that the pair frequencies marginalize to the site-specific frequencies and that the site-specific frequencies are normalized to one, that is, the following conditions have to be fulfilled

$$\sum_a P_{ij}(a,b) = P_j(b) \quad \forall i<j;\forall b \tag{3.19a}$$

$$\sum_b P_{ij}(a,b) = P_i(a) \quad \forall i<j;\forall a \tag{3.19b}$$

$$\sum_a P_i(a) = 1 \quad \forall i \tag{3.19c}$$

By substituting the pair frequencies by the correlation measure Q, the conditions can be rewritten to

$$\sum_a P_i(a)Q_{ij}(a,b) = 1 \quad \forall i<j;\forall b \tag{3.20a}$$

$$\sum_b P_j(b)Q_{ij}(a,b) = 1 \quad \forall i<j;\forall a \tag{3.20b}$$

Thus, the function that has to be minimized now reads in the cluster expansion as

$$\begin{aligned}
&F_2^{(\text{cluster})}(\{Q_{ij}(a,b)\},\{P_i(a)\};\Lambda_E,\Lambda_e,\eta_{ij}(a),\zeta_{ij}(b),\zeta_i)\\
=&S_2(\{Q_{ij}(a,b)\},\{P_i(a)\})\\
&-\Lambda_E\left(\langle E_{\text{nat}}\rangle_{\text{evol}}(\{Q_{ij}(a,b)\},\{P_i(a)\})-\langle E_{\text{nat}}\rangle^0_{\text{evol}}\right)\\
&-\Lambda_e\left(\left\langle E^{(\text{REM})}_{\text{min}}\right\rangle_{\text{evol}}(\{Q_{ij}(a,b)\},\{P_i(a)\})-\left\langle E^{(\text{REM})}_{\text{min}}\right\rangle^0_{\text{evol}}\right)\\
&+\sum_{i<j}\sum_b\left[\eta_{ij}(b)\left(\sum_a[P_i(a)Q_{ij}(a,b)]-1\right)\right]+\sum_{i<j}\sum_b\left[\zeta_{ij}(a)\left(\sum_b[P_j(b)Q_{ij}(a,b)]-1\right)\right]\\
&+\sum_i\theta_i\left[\sum_a[P_i(a)]-1\right] \quad .
\end{aligned} \tag{3.21}$$

The Lagrange parameters $\eta_{ij}(a)$ and $\zeta_{ij}(b)$ regulate the constraints on the correct marginalization of the pair-specific frequencies and the Lagrange parameter θ_i the correct normalization of the site-specific frequencies.

Now, it is desired to minimize the function $F_2^{(\text{cluster})}$ with respect to the site-specific frequencies and the correlation measure Q. In principle, the site-specific frequencies can be inferred together with the correlation measure Q in the same framework. However, I will restrict myself to the minimization with respect to the correlation measure, because site-specific frequencies were inferred successfully in previous studies as discussed in the introduction. Nevertheless, it is interesting to compare site-specific probabilities obtained from the two schemes. Since the derivation of site-specific probabilities is very demanding if inferred together with Q, one could infer them from the cluster expansion of single sites to simplify the scheme.

At the point of minimal free energy the derivative of the function $F_2^{(\text{cluster})}$ (eq. (3.21)) with respect to $Q_{ij}(a,b)$ vanishes,

$$\begin{aligned}\frac{\partial F_2^{(\text{cluster})}}{\partial Q_{ij}(a,b)} =& P_i(a)P_j(b)\Big[-1 - \ln Q_{ij}(a,b) - \Lambda_E C_{ij}^{\text{nat}} U(a,b) - \Lambda_e \left\langle C_{ij} \right\rangle U(a,b) \\ &+ \Lambda_e U^2(a,b) \frac{1}{2} \langle \sigma(E) \rangle_{\text{evol}}^{-1} \sigma_{ij} \sqrt{2 \ln N_{\text{mis}}} - \eta_{ij}(a)/P_j(b) - \zeta_{ij}(b)/P_i(a) \Big] \stackrel{!}{=} 0 \quad .\end{aligned} \tag{3.22}$$

By rearranging the equation and introducing the abbreviation $R = \frac{1}{2} \langle \sigma(E) \rangle_{\text{evol}}^{-1} \sqrt{2 \ln N_{\text{mis}}}$, which has a positive sign, the following expression for Q is found,

$$Q_{ij}(a,b) = \theta_{ij}(a)\xi_{ij}(b) \exp\left(-(\Lambda_E C_{ij}^{\text{nat}} + \Lambda_e \left\langle C_{ij} \right\rangle) U(a,b) + \sigma_{ij} R \Lambda_e U^2(a,b) \right) \quad . \tag{3.23}$$

The Lagrange multipliers $\theta_{ij}(a)$ and $\xi_{ij}(b)$ are redefinitions of $\eta_{ij}(a)$ and $\zeta_{ij}(b)$ in eq. (3.22) and have a positive sign. Note that the equation (3.23) is still implicit for Q as R depends on Q via $\langle \sigma(E) \rangle_{\text{evol}}$. Nevertheless, from eq. (3.23) one can guess the sign of the Lagrange multipliers Λ_E and Λ_e. For a large negative free energy of the native state, a pair of amino acids with a strongly attractive interaction, i.e., with a negative interaction $U(a,b)$, is expected to be positively correlated. Since the exponent is $-\Lambda_E C_{ij}^{\text{nat}} U(a,b)$, Λ_E is expected to be positive. The minimal energy of the misfolded state becomes larger, as the mean free energy of the misfolded ensemble $\langle E \rangle$ grows positive, that is, pairs of amino acids with a repulsive interaction are expected to be correlated. The second centered moment of misfolded ensemble becomes smaller if pairs with a strong interaction, positive or negative, are avoided. From the exponent it is easy to see that both criteria are met if Λ_e is negative.

For a prediction of Q, the Lagrange multipliers $\theta_{ij}(a)$ and $\xi_{ij}(b)$ are still to be determined. In principle, this can be achieved by starting from an initial guess of Λ_E, Λ_e and R and then determining $\theta_{ij}(a)$, $\xi_{ij}(b)$ numerically from the constraints (3.19c)-(3.20b) and afterwards adjusting Λ_E, Λ_e and R towards the correct constraints of the free energies $\left\langle E_{\text{nat}} \right\rangle_{\text{evol}}^0$ and $\left\langle E_{\text{min}} \right\rangle_{\text{evol}}^0$.

However, to attain a better analytical insight and more simple equations, correlations are assumed to be weak, that is, $Q_{ij}(a,b)$ is close to one. Then, the first order Taylor expansion about one of the logarithm in eq. (3.22), i.e., $\ln Q_{ij}(a,b) \approx Q_{ij}(a,b) - 1$, is justified and the linear equation is found,

$$Q_{ij}(a,b) = -\Lambda_E C_{ij}^{\text{nat}} U(a,b) - \Lambda_e \left(U(a,b) \left\langle C_{ij} \right\rangle - U^2(a,b) \sigma_{ij} R \right) - \eta'_{ij}(a) - \zeta'_{ij}(b) \quad , \tag{3.24}$$

where I introduced the abbreviations $\eta'_{ij}(a) = \eta_{ij}(a)/P_j(b)$ and $\zeta'_{ij}(b) = \zeta_{ij}(b)/P_i(a)$. The linearized equation allows to get rid of the Lagrange multipliers η' and ζ' that control the marginalization of $Q_{ij}(a,b)$. To this end, eq. (3.24) is substituted into the marginalization conditions eq. (3.20a) and eq. (3.20b), thus obtaining

$$-\zeta'_{ij}(b) + \sum_a P_i(a) \left[-\eta'_{ij}(a) - \Lambda_E C_{ij}^{\text{nat}} U(a,b) - \Lambda_e \left(\langle C_{ij} \rangle U(a,b) - R \sigma_{ij} U^2(a,b) \right) \right] = 1 \tag{3.25a}$$

$$-\eta'_{ij}(a) + \sum_b P_j(b) \left[-\zeta'_{ij}(b) - \Lambda_E C_{ij}^{\text{nat}} U(a,b) - \Lambda_e \left(\langle C_{ij} \rangle U(a,b) - R \sigma_{ij} U^2(a,b) \right) \right] = 1 \quad . \tag{3.25b}$$

These equations still couple the Lagrange parameters ζ' and η'. However, by substituting the Lagrange parameters in eq. (3.24) by the first occurrences of the Lagrange parameters in eq. (3.25a) and eq. (3.25b), one sees that the problem reduces to the determination of only one parameter instead of forty, since the terms $\sum_a P_i(a)\,\eta'_{ij}(a)$ and $\sum_b P_j(b)\,\zeta'_{ij}(b)$ depend neither on a or b. Thus, one can write

$$\begin{aligned} Q_{ij}(a,b) =& 2 - \sum_{a'} P_i(a') \left[-\eta'_{ij}(a') - \Lambda_E\, C_{ij}^{\text{nat}}\, U(a',b) - \Lambda_e \left(\langle C_{ij} \rangle U(a',b) - R\,\sigma_{ij}\, U^2(a',b) \right) \right] \\ & - \sum_{b'} P_j(b') \left[-\zeta'_{ij}(b') - \Lambda_E\, C_{ij}^{\text{nat}}\, U(a,b') - \Lambda_e \left(\langle C_{ij} \rangle U(a,b') - R\,\sigma_{ij}\, U^2(a,b') \right) \right] \\ & - \Lambda_E\, C_{ij}^{\text{nat}}\, U(a,b) - \Lambda_e \left(\langle C_{ij} \rangle U(a,b) - R\,\sigma_{ij}\, U^2(a,b) \right) \quad . \end{aligned} \tag{3.26}$$

By sorting by the parameters Λ_E and Λ_e I obtain

$$\begin{aligned} Q_{ij}(a,b) =& C - (\Lambda_E\, C_{ij}^{\text{nat}} + \Lambda_e) \left(U(a,b) - \sum_{a'} P_i(a')\, U(a',b) - \sum_{b'} P_j(b')\, U(a,b') \right) \\ & + \Lambda_e R\,\sigma_{ij} \left(U^2(a,b) - \sum_{a'} P_i(a')\, U^2(a',b) - \sum_{b'} P_j(b')\, U^2(a,b') \right) \quad , \end{aligned} \tag{3.27}$$

with $C \equiv 2 + \sum_{a'} P_i(a')\,\eta'_{ij}(a') + \sum_{b'} P_j(b')\,\zeta'_j(b')$. Now the constant term C is computed by substituting $Q_{ij}(a,b)$ from eq. (3.27) into the marginalization conditions again yielding the following equation for all forty constraints

$$\begin{aligned} C - \Lambda_E\, C_{ij}^{\text{nat}} \sum_{a',b'} P_i(a')\, P_j(b')\, U(a',b') - \Lambda_e \langle C_{ij} \rangle \sum_{a',b'} P_i(a')\, P_j(b')\, U(a',b') & \\ + \Lambda_e R\,\sigma_{ij} \sum_{a',b'} P_i(a')\, P_j(b')\, U^2(a',b') = 1 \quad . & \end{aligned} \tag{3.28}$$

Plugging C back into eq. (3.27), an equation for $Q_{ij}(a,b)$ is gained, which depends only on the Lagrange parameters of the stability constraints

$$Q_{ij}(a,b) = 1 - \left(\Lambda_E\, C_{ij}^{\text{nat}} + \Lambda_e \langle C_{ij} \rangle \right) F_{ij}^{(1)}(a,b) + \Lambda_e R\,\sigma_{ij} F_{ij}^{(2)}(a,b) \quad , \tag{3.29}$$

where I introduced the abbreviations

$$F_{ij}^{(1)}(a,b) = U(a,b) - \sum_{a'} P_i(a')\, U(a',b) - \sum_{b'} P_j(b')\, U(a,b') + \sum_{a',b'} P_i(a')\, P_j(b')\, U(a',b') \tag{3.30a}$$

$$F_{ij}^{(2)}(a,b) = U^2(a,b) - \sum_{a'} P_i(a')\, U^2(a',b) - \sum_{b'} P_j(b')\, U^2(a,b') + \sum_{a',b'} P_i(a')\, P_j(b')\, U^2(a',b') \quad . \tag{3.30b}$$

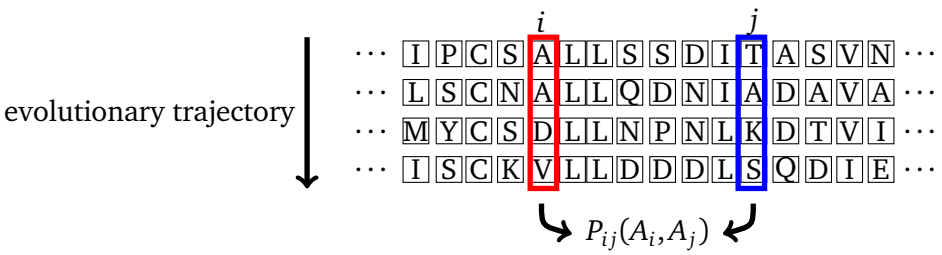

Figure 3.1: Measuring pair frequencies from an alignment.

3.2.3 Application to alignments

Evolutionary information, like correlated mutations, are contained in multiple sequence alignments. Often the same protein can be found in different species and populations, which have a different amino acid sequence. In an MSA sequences of a protein from different species are arranged such that equivalent sites are identified with each other. The alignment can be represented as a matrix, where in each row is one sequence and each column represents a site in the protein (Fig. 3.1).

The probability $P_{ij}(a, b)$ of two amino acids a and b, which can be found at site i and j respectively, can be measured by counting the amino acids from MSA (Fig. 3.1). However, statistical analysis of MSA data is complicated. Alignment data suffers from sampling biases and correlations that stem from common evolution rather than selection forces, which need complex mathematical models to be purged from the data [49, 50, 63]. Additionally, MSA data does not provide enough counts for reliable values of Q, i.e., large statistical fluctuations are expected. All these problems render a detailed quantitative analysis with sequence alignments difficult if not impossible. Therefore, in an alternative approach correlated substitutions are inferred by a statistical analysis of protein structures and sequences from the PDB, which are also a product of evolution. This is done by grouping together pairs of residues that are considered structurally equivalent. Such an analysis can be understood as a sequence alignment, where the sequence of residues and the affiliation of residues to a specific protein are neglected.

An alternative is the simulation of sequences, which can produce an arbitrarily large number of sequences and is free from sampling biases. In addition, simulations allow for explicitly imposing conditions on the thermodynamic stability as formulated in our model, which makes the assessment of the model particularly meaningful.

3.3 Results

3.3.1 Simulation data

In this Section I describe the production of protein sequences by simulations, from which I measure the correlation of amino acids at two different sites. In principle, the parameters regulating the energetic constraints can be inferred from simulation data in two ways. First, the parameters are obtained by fitting the predicted correlations to the ones observed in the simulation. Second, the parameters are determined such that the theory reproduces the evolutionary averages of the

energies, which are measured in the simulations and used as constraints in derivation of the theory. The quality of the prediction of correlated substitutions of the theory is then assessed.

Simulation of structurally constrained protein sequence evolution

In previous studies simulations of protein sequence evolution that explicitly account for thermodynamic stability model were performed [64]. Here, neutral evolution is simulated, which assumes that mutations are either selective neutral, that is to say, viable sequences have equal fitness, or lethal, which results in a fitness of zero. In accordance with earlier studies I impose constraints on the folding stability by restricting the free energy of the native state E_{nat} and the energy gap α,

$$\alpha \equiv \min_{C} \frac{E_{\text{nat}} - E(C,A)}{E_{\text{nat}}(1 - q(C^{\text{nat}}, C))} \tag{3.31}$$

which measures the stability against misfolding and ensures a well correlated energy landscape. In principle, the energy gap can be measured by *threading* (see Section 2.2.2), but here, in order to speed up computations, the energy gap is estimated. Since the contact overlap $q(C^{\text{nat}}, C)$ is narrowly distributed around a small value for most misfolded structures C, it is set to a typical value $q(C^{\text{nat}}, C) = q_0 \equiv 0.1$.

Again, the minimal energy of the misfolded ensemble E_{min} is approximated with the REM. To simplify matters, the contact frequency is assumed to be independent of the distance of the residues in sequence and the number of contacts of a misfolded structure is estimated with the number of contacts N_C of the native state. Thus, the mean and standard deviation of the misfolded ensemble becomes

$$\langle E \rangle \approx N_C \left[U\right] \tag{3.32a}$$

$$\sqrt{\left\langle (E - \langle E \rangle)^2 \right\rangle} \approx \left(\left[U^2\right] - \left[U\right]^2 \right)^{\frac{1}{2}} \sqrt{N_C} \quad . \tag{3.32b}$$

Thus, the energy gap becomes

$$\alpha \approx \frac{E_{\text{nat}} - E_{\text{min}}}{q_0 E_{\text{nat}}} \tag{3.33}$$

with

$$E_{\text{min}} \approx N_C \left[U\right] - \left(\left[U^2\right] - \left[U\right]^2 \right)^{\frac{1}{2}} \sqrt{2 N_C \log N_{\text{mis}}} \tag{3.34a}$$

$$\log N_{\text{mis}} \equiv 4 + 0.1\, L \quad . \tag{3.34b}$$

A sequence is considered viable, i.e., it has fitness of one, if the free energy of the native state E_{nat} is smaller than the threshold $E_{\text{nat}}^{\text{thr}}$ and the energy gap α is larger than the threshold α_{thr}. Following previous studies, the thresholds are fixed to values such that the native sequence found together with the native structure in the PDB file is marginally viable, that is, I set $E_{\text{nat}}^{\text{thr}}$ to $0.98 \times E_{\text{nat}}(A_{\text{nat}})$ and α_{thr} to $0.98 \times \alpha(A_{\text{nat}})$.

The sequence is represented by the DNA sequence, which is mapped on the amino acid sequence using the standard genetic code (see Table A.2 on page 93). In each step one randomly

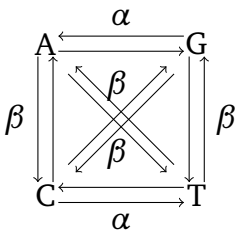

Figure 3.2: Two parameter mutation model. Each nucleotide can mutate into one of the other three. The rates are assumed to differ between transition α and transversions β, here a common rate ration of $\alpha/\beta = 2$ is assumed.

selected nucleotide is mutated to one of the other three possible nucleotides with probabilities from a mutation model, where the rate between transition and transversion is set to two (Fig. 3.2). If the mutation changes the coded amino acid, i.e., the mutation is not synonymous, the free energy and the energy gap are recomputed and the fitness is determined for the new values. If the fitness is zero, the mutation is rejected and the weight w_n of the current amino acid sequence is increased by one. Otherwise, the mutation is accepted and the amino acid is replaced by the new one. If the mutation is synonymous, i.e., the amino acid sequences does not change, the mutation is accepted and the weight of the current amino acid sequence is increased by one.

The probability of finding an amino acid a at site i and an amino acid b at j is computed by adding the weight of the sequences, in which the two amino acids occur, and afterwards dividing by the weight of all sequences. The site-specific probability is determined in an analogous manner. With the Kronecker-delta $\delta(a,b)$, which is one if $a = b$ and zero otherwise, the probabilities can be written as

$$P_{ij}(a,b) = \frac{\sum_n w_n \delta(a, A_i^{(n)}) \delta(b, A_j^{(n)})}{\sum_n w_n} \tag{3.35a}$$

$$P_i(a) = \frac{\sum_n w_n \delta(a, A_i^{(n)})}{\sum_n w_n} \quad . \tag{3.35b}$$

The computation of the correlation measure Q is then straightforward by substituting the a posteriori probabilities into the definition of the correlation measure (eq. (3.1)).

This procedure gives rise to a trajectory through sequence space, where two successive sequences differ by one amino acid. A pair of amino acids at two sites changes if one of the sites changes. Assuming that each site is substituted equally often, the probability that a pair changes in one step is equal to the probability that one of the sites is substituted, which is as small as $2/L$. That is to say, in most of the steps a pair is not changed and even for a long trajectory a small number of different pairs can be expected, which can give rise to biases in the correlation and strong statistical fluctuations. To test for such biases due to statistics and common evolution, I ran a simulation without imposing constraints on the thermodynamic stability. I computed the mutual information (eq. (3.2)) and the average Q over all pairs and amino acids. For 5×10^6 random substitutions the average MI of two positions is as low as 10^{-11} and the average Q is as low as 1 ± 0.02. Thus, correlations due to common evolution and statistical fluctuations are very small and can be neglected in the following. However, the statistical fluctuations are expected to

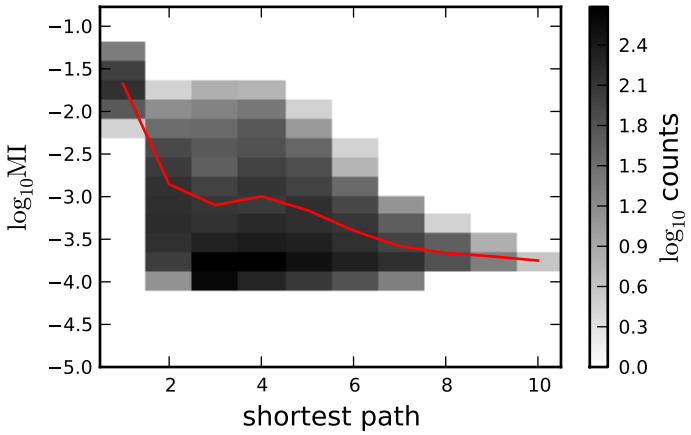

Figure 3.3: The logarithm of mutual information measured from simulations versus the length of shortest path in the contact network of the protein 3rn3. The plot shows a two dimensional histogram, counts are color coded in logarithmic scaling, and the red line indicates the mean of the logarithm for a certain path length.

be larger if simulation is done with constraints, because the single site frequencies are strongly skewed, while in the simulation without constraints the background probability distribution is attained. However, even if one would constrain only the single site frequencies to their values from the constrained simulation, the mutual information is expected to be very small.

I consider four different small globular proteins with the PDB-identifiers 1iro, 1ubq, 3rn3, and 451c of different secondary structure composition and folding topology. As the results are qualitatively the same for all four chains, they are exemplified by the chain 3rnr3.

Correlation and relative position in contact network

Before testing the model it is worthwhile to analyze the amount of correlation of different pairs of residues. In particular, the model expanded to the second order of the cluster expansion (eq. (3.23)) suggests that the correlations depend on whether two contacts are in contact or not and have a small modulation due to the marginalization constraints on the site-specific frequencies. However, regarding the relative position within the contact network, the simulated data shows a more intricate relationship.

As measure for the relative position I use the length of the shortest path which connects two residues. Fig. 3.3 shows the relation between mutual information and the length of the shortest path. Residues which are in contact, i.e., which are connected by a path of length one, are strongly correlated and exhibit the largest mutual information, which decreases rapidly with increasing path length. In particular, the larger the distance the smaller is the largest mutual information observed. However, the variation of the mutual information is not completely explained by the length of the shortest path as it varies for residues in the same distance over many orders of magnitude.

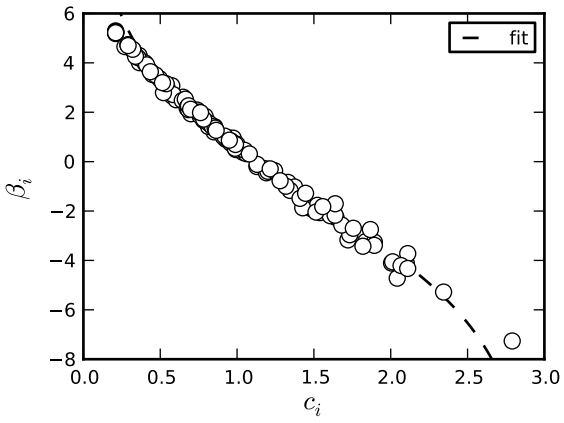

Figure 3.4: Fitted parameters β_i to all site-specific distributions of the protein with PDB-ID 3rn3. For each site i eq. (1.11) is fitted to the site-specific distribution, obtaining one parameter β_i for each site. This parameter is strongly related to the effective connectivity c_i, in the sense that the EC is well correlated with the evolutionary averaged hydrophobicity described by eq. (3.36). The dotted line shows a fit of the evolutionary averaged hydrophobicity obtained by the fitted parameter β_i to the profile via eq. (3.37).

Fitting to data

Before fitting the model to pair correlations I repeat the analysis from previous studies and fit the model to site-specific probabilities. In particular, for each site i the parameter β_i in eq. (1.11) is fitted to the observed amino acid distribution at each site i, yielding a very high correlation coefficient of the fit (0.97 on average), that is, the site-specific probability can be very well described as a Boltzmann distribution of temperature β_i. The temperature can be computed from the value of the EC at site i. The EC is assumed to be perfectly correlated with the mean hydrophobicity $\left[h\right]$, which can be computed from the Boltzmann distribution by combining eq. (1.8) and eq. (1.11),

$$\left[h\right](\beta) = \sum_a h(a) \frac{n_c(a)}{Z(\beta)} \exp(-\beta\, h(a)) \quad . \tag{3.36}$$

A perfect correlation between two vectors requires to map one on the other by a transformation of scale and shift,

$$c_i = A\left[h\right](\beta_i^{\text{opt}}) + B \quad . \tag{3.37}$$

The parameters A and B correspond to the ratio of the variances of the two vectors and the difference in their means, respectively, and are obtained from a fit to the fitted β_i and the EC-value c_i. Fig. 3.4 shows that the temperature β_i is strongly related to the EC. That is, site-specific frequencies can be very well predicted for simulated data from only two parameters A and B. However, for the analysis of correlations measured instead of predicted single site frequencies are used to ensure that the predicted and measured correlation measures Q obey the same marginalization.

	Λ_E	Λ_e	Λ_{e2}
1iroA	3.67	-0.525	-0.814
1ubqA	2.70	-0.209	-0.329
3rn3A	1.99	-0.104	-0.169
451cA	2.52	-0.189	-0.313

Table 3.1: Fitted Lagrange parameters to simulated sequence evolution of five different proteins. The parameters are obtained by fitting eq. (3.38) to measured correlations $Q_{ij}^{\text{meas}}(a,b)$.

When the correlations are predicted from (3.29), the problem occurs that the theory still contains the term R that depends on the correlation measure Q and renders eq. (3.29) implicit, which is much more demanding to solve than a linear equation. In principle, the term R could be measured from simulations. Instead, eq. (3.29) is made explicit by regarding the term R as a fitting parameter, which is determined in the following in the same manner as the other Lagrange parameters. In order to simplify notation, I introduce a third parameter $\Lambda_{e2} = R\Lambda_e$, which is considered independent of the other two Lagrange parameters. Indeed, the introduction of Λ_{e2} produces a theory which one would have obtained if the mean and standard deviation of the energy of the misfolded ensemble would were constrained separately, that is, the Lagrange parameter Λ_e regulates the constraint on $\langle\langle E\rangle\rangle_{\text{evol}}$, while Λ_{e2} regulates the constraint on $\langle\sigma(E)\rangle_{\text{evol}}$.

In the simulation the minimal energy is estimated with the REM, that is, the mean contact frequency $\left\langle C_{ij}\right\rangle$ and σ_{ij} are assumed to be independent of the indices i and j. Since the simulations rely on a more simple model, it is advisable to simplify the theory in the same manner. As a consequence, the contact frequencies and contact correlations can be absorbed in the Lagrange parameters Λ_e and Λ_{e2}, respectively, which then results in the fitting formula

$$Q_{ij}^{(\text{pred})}(a,b) = 1 - (\Lambda_E C_{ij}^{\text{nat}} + \Lambda_e)F_{ij}^{(1)}(a,b) + \Lambda_{e2}F_{ij}^{(2)}(a,b)) \quad . \tag{3.38}$$

Now, the three Lagrange parameters are determined by a least square fit of eq. (3.38) of all pairs $\{(i,j)|0 \le i < j-2 < L\}$ of sites. Apparently, the contact matrix C_{ij}^{nat} in the fitted function ensures that Λ_E is determined by all pairs in contact, whereas Λ_e is determined only by pairs which are not in contact. The parameter Λ_{e2} is determined by both classes of pairs.

The values of the fitted parameters are listed in Table 3.1 for the five test proteins. The values of the parameters are in accordance with our expectations: Λ_E is large and positive, reflecting that direct contacts are strongly correlated. Λ_e and Λ_{e2}, which represent together the parameter Λ_e in eq. (3.29), are negative and have a smaller absolute value.

The quality of the fit is assessed by three measures: First, the relative error is considered, which is defined as the mean square deviation of all 400 values of Q for one pair of residues divided by the variances of the measured Q. The division by the variance permits a meaningful comparison of the fitting error of pairs with a different strength of correlation, i.e., different magnitudes of deviations of Q from one. The second measure is the correlation coefficient of the 400 values of each pair i j. Third, it is assessed whether the scale correlation is correctly predicted by the means of the ratio of the standard deviation of the predicted and measured Q.

As the correlation seems to vary with the distance in the contact network, the assessment is performed for every minimal distance in the contact network separately. As an example, I discuss

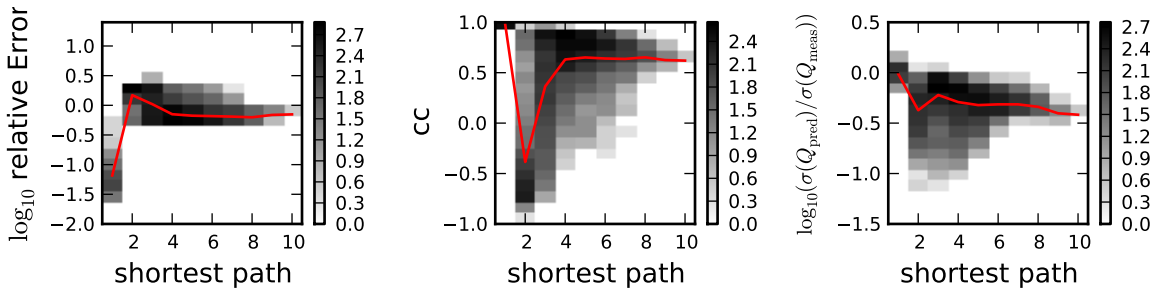

Figure 3.5: Assessment of fitting correlated substitutions theory to simulated data. Relative error, correlation coefficient and ratio of scales between measured and predicted correlation measures versus length of shortest path in contact network.

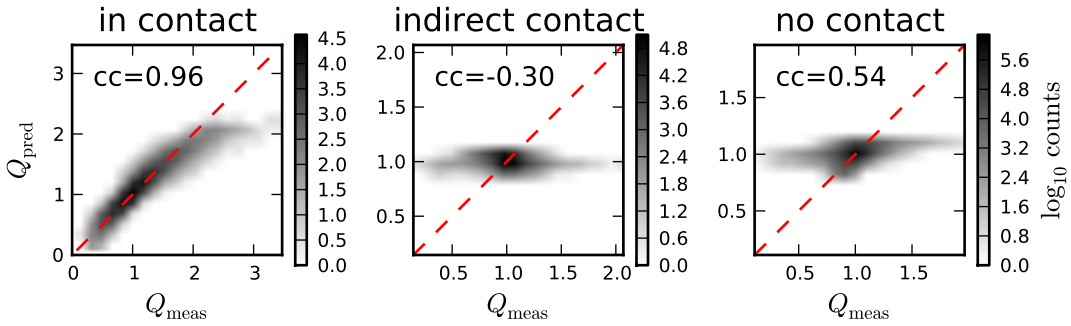

Figure 3.6: Density plot of predicted versus measured Q for protein 3rn3 for direct, indirect contacts, and pairs which have neither a direct nor an indirect contact.

the quality for the protein 3rn3; the results of the other proteins look alike. The distribution of the three measures is shown as a density plot in Fig. 3.5. For direct contact the correlation is predicted very well, the average correlation coefficient is 0.96 and the relative error is very small (see Fig. 3.6). For indirect contacts the quality of the prediction becomes very low; at that point the correlation coefficient reaches its minimum and becomes even negative on average and the relative error reaches here its maximum. For larger distances the quality of the fit increases again: The correlation coefficient levels off at a value of approximately 0.55, in particular, the minimal correlation coefficient increases. Similarly, the relative error decreases again, however, not to the values reached for pairs in contact. The variance of the correlation measure is only correctly estimated for pairs in contact. This is to no surprise since for larger distances the fit cannot reproduce the correlation pattern (low average cc) of non-contacts. Therefore the fit reduces the variance of the prediction when minimizing the error.

Apparently, for pairs in contact the direct interaction dominates the correlation and is very well predicted by the theory here, which considers only correlation induced by the direct interaction of two residues. However, for pairs, which are not in contact, indirect interactions via a third residue or even a network, established by contacts between the two residues, have an equally important contribution and cannot be predicted by means of the theory presented here. However, at least for indirect contacts the contribution of indirect correlations can be estimated and improves the prediction as discussed in the next Section.

Eq. (3.29) does not couple different pairs, but an indirect correlation of two residues is expected if they are strongly correlated with a third residue. Such indirect correlations can travel even longer distances through the contact network, as suggested by Fig. 3.3. In principle, one can incorporate in our ansatz the correlation of three residues by including the third order term in the cluster expansion, which considers the probability $P_{ijk}(a,b,c)$ of all triples i, j, k of residues. However, the analytical treatment is rather complicated. Instead, I make an *ad hoc* ansatz which yields a three body correction to the predicted correlations $Q_{ij}^{\text{pred}}(a,b)$.

I consider three random variables a, b and c, where a and c represent residues forming an indirect contact via b. Since the correlation between residues in contact is much stronger than for non-contacts, it is reasonable to assume that b depends only on a and c depends only on b. These dependencies are given by the conditional probabilities $P(b|a)$ and $P(c|b)$. Then, the probability $P(a,b,c)$ can be written as $P(a,b,c) = P(a)P(b|a)P(c|b)$. By marginalizing over b, one finds the pair probability of a and c. The indirect correlation $Q(a,b)$ is found by dividing by the product $P(a)P(c)$,

$$\begin{aligned} Q(a,c) = \frac{P(a,c)}{P(a)P(c)} &= \frac{\sum_b P(a)P(b|a)P(c|b)}{P(a)P(c)} \\ &= \sum_b \frac{P(b,a)P(c,b)}{P(a)P(c)P(b)} = \sum_b Q(a,b)\,Q(b,c)\,P(b) \quad . \end{aligned} \tag{3.39}$$

This ansatz yields the indirect correlation established by one indirect contact. However, in many cases indirect contacts are established by more than one residue. A derivation of the indirect correlations for more than one residue is not straightforward. Therefore, I simply define the correction to the predicted correlation as the sum over all residues establishing an indirect contact. Then, I add the contribution due to indirect contacts ΔQ to the direct correlation predicted by the theory (eq. (3.38)),

$$\Delta Q_{ij}(a,b) = (1 - C_{ij}^{\text{nat}}) \sum_{k \neq i,j} C_{ik}^{\text{nat}}\, C_{jk}^{\text{nat}} \left[\sum_c \left(Q_{ik}^{\text{pred}}(a,c)\, Q_{jk}^{\text{pred}}(b,c)\, P_k(c) \right) - 1 \right] \quad . \tag{3.40}$$

Note that $Q_{ij}^{\text{pred}}(a,b) + \Delta Q_{ij}(a,b)$ does fulfill the marginalization conditions eqs. (3.20), as the contribution of the correction term to the marginalizing sums is zero.

The impact of the correction term on the quality of the prediction is shown in Fig. 3.7. Indeed, the average correlation coefficient is significantly improved from -0.38 to 0.62 and the relative error is reduced, with the exception of a few residues, for which it is increased.

Indeed, a proper treatment of indirect correlation is computational feasible, however, very demanding. In a recent study Weigt *et al.* have formulated a theory for correlated mutations that is similar to the theory developed here. From a maximum entropy argument under the constraints of marginalization to pair and site specific frequencies, they found the probability distribution of the sequence [53],

$$P_{1,\ldots,L}(A_1,\ldots,A_N) = \frac{1}{Z} \exp\left(-\sum_{i<j} \eta_{ij}(A_i,A_j) + \sum_i f_i(A_i) \right) \quad . \tag{3.41}$$

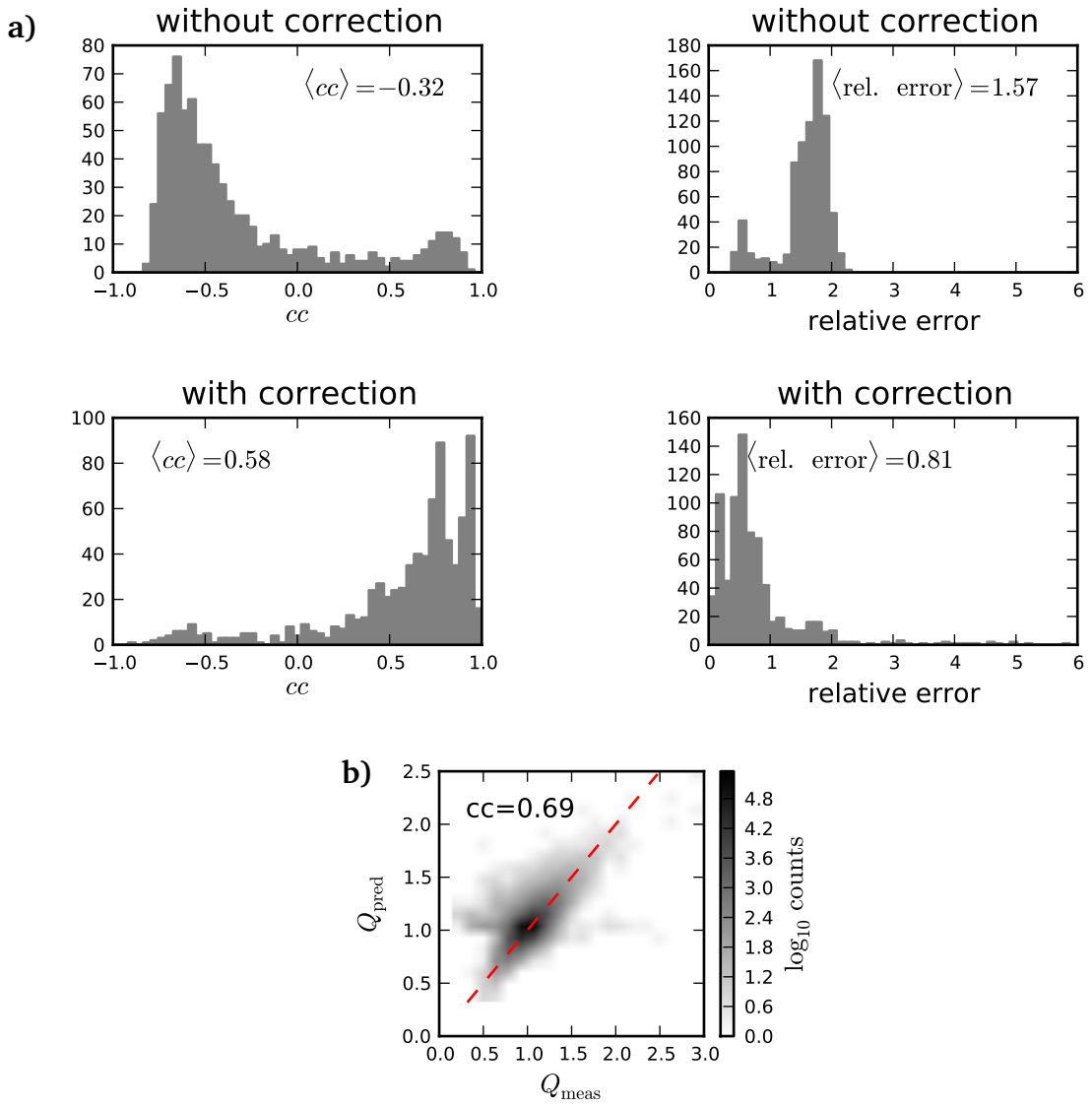

Figure 3.7: Assessment of the correction due to indirect correlations. **a)** Histograms of the correlation coefficient and relative error of the prediction with and without the correction term. Both the correlation coefficient and the relative error improve on average under the application of the correction term. The relative error, however, increases for several pairs (relative error > 2). **b)** Density plot of Q^{meas} versus Q^{pred} after the correction.

The parameters $\eta_{i,j}(A_i, A_j)$ and $f_i(A_i)$ are fitted, such that the marginalized pair specific and site specific probabilities match the to corresponding frequencies measured from the MSA. This requires the marginalization of the probability eq. (3.41) to pair-specific probabilities, which is done with belief propagation [53, 65]. Such a computationally involved scheme is in principle applicable to our problem as well.

Furthermore, comparing eq. (3.41) to eq. (3.11), one finds that the parameters η_{ij} correspond to the direct interaction between the residues $\Lambda_E U(A_i,A_j) - \Lambda_e U(A_i,A_j) + \Lambda_{e2} U^2(A_i,A_j)$. Thus,

from point of view of our model, Weigt *et al.* inferred the direct interaction between residues. Then, they defined the pair specific probability, which is due to direct interactions only,

$$P_{ij}(A_i,A_j) = \frac{1}{Z_{ij}} \exp\left(\eta_{ij}(A_i,A_j) + f_i(A_i) + f_j(A_j)\right) \quad . \tag{3.42}$$

This direct probability is different from the observed pair frequency in the MSA data, which arose from direct and indirect interactions. Indeed, this direct probability corresponds to pair probability after the the cluster expansion (cf. (3.23)). This analogy shows that indirect correlations are difficult to handle.

Obtaining Lagrange parameters from constraints

Alternatively to fitting, the three Lagrange parameters can be obtained by adjusting them such that the conditions on the evolutionary average free energy E_{nat} and $E_{\min}$ are met. One would have obtained eq. (3.38) if constraints on $\langle [U] \rangle_{\text{evol}}$ and $\left\langle [U^2] \right\rangle_{\text{evol}}$ instead of $\langle E_{\min} \rangle_{\text{evol}}$ are imposed, which is exploited here to determine the three Lagrange parameters from the evolutionary averages measured from the simulation. To this end, the terms $Q_{ij}(a,b)$ in $\langle E_{\text{nat}} \rangle_{\text{evol}}$ (eq. (3.5)) and in the evolutionary averages of $[U^i] = 1/N_p \sum_{i<j-2} P_i(a) P_j(b) Q_{ij}(a,b) U^i(a,b)$ and $[U^2]$ are substituted by the defintion of $Q_{ij}^{\text{pred}}(a,b)$ from eq. (3.38). By rearranging the terms one obtains expressions for the evolutionary averaged energies as a linear function of the Lagrange parameters,

$$\langle E_{\text{nat}} \rangle_{\text{evol}} = \sum_{i<j-2} \sum_{a,b} C_{ij}^{\text{nat}} U(a,b) P_i(a) P_j(b) Q_{ij}^{\text{pred}}(a,b) \quad = c_{00} + c_{01}\Lambda_E + \Lambda_e c_{02} + \Lambda_{e2} c_{03} \tag{3.43a}$$

$$\langle [U] \rangle_{\text{evol}} = \frac{1}{N_p} \sum_{i<j-2} \sum_{a,b} P_i(a) P_j(b) U(a,b) Q_{ij}^{\text{pred}}(a,b) \quad = c_{10} + c_{11}\Lambda_E + \Lambda_e c_{12} + \Lambda_{e2} c_{13} \tag{3.43b}$$

$$\left\langle [U^2] \right\rangle_{\text{evol}} = \frac{1}{N_p} \sum_{i<j-2} \sum_{a,b} P_i(a) P_j(b) U^2(a,b) Q_{ij}^{\text{pred}}(a,b) \quad = c_{20} + c_{21}\Lambda_E + \Lambda_e c_{22} + \Lambda_{e2} c_{23} \tag{3.43c}$$

with the coefficients

$$\begin{aligned} c_{m0} &= \sum_{i<j} \sum_{a,b} X_{ij}^{(m)}(a,b) P_i(a) P_j(b) & c_{m1} &= -\sum_{i<j} \sum_{a,b} C_{ij}^{\text{nat}} X_{ij}^{(m)}(a,b) P_i(a) P_j(b) F_{ij}^{(1)}(a,b) \\ c_{m2} &= -\sum_{i<j} \sum_{a,b} X_{ij}^{(m)}(a,b) P_i(a) P_j(b) F_{ij}^{(1)}(a,b) & c_{m3} &= \sum_{i<j} \sum_{a,b} X_{ij}^{(m)}(a,b) P_i(a) P_j(b) F_{ij}^{(2)}(a,b) \end{aligned} \tag{3.44}$$

where

$$X_{ij}^{(0)}(a,b) = C_{ij}^{\text{nat}} U(a,b) \qquad X_{ij}^{(1)}(a,b) = \frac{1}{N_p} U(a,b) \qquad X_{ij}^{(2)}(a,b) = \frac{1}{N_p} U^2(a,b) \quad . \tag{3.45}$$

constraint		c_{i0}	c_{i1}	c_{i2}	c_{i3}
E_{nat}	=-22.76	-1.56e+01	-3.84e+00	-3.84e+00	-1.22e+00
$\left\langle \left[U^2\right] \right\rangle_{\mathrm{evol}}$	=0.0258	+2.61e-02	+1.60e-04	+2.39e-03	+1.46e-03
$\left\langle \left[U\right] \right\rangle_{\mathrm{evol}}$	=0.00469	+3.93e-03	-5.04e-04	-9.86e-03	-2.39e-03

Table 3.2: Coefficients of linear function of evolutionary averaged energies from Lagrange parameters for protein chain 3rn3A. The left column shows the values of the evolutionary averaged E_{nat}, $\left[U\right]$ and $\left[U^2\right]$ as measured from simulations. The other columns list the coefficients to the Lagrange parameters in eq. (3.43), which are computed from the site-specific probabilities measured from simulations. The coefficients c_{i0} are constants in the linear function and represent the evolutionary averaged energies without correlated substitutions, i.e., $Q = 1$.

a)	Λ_E	Λ_e	Λ_{e2}
1iroA	3.55	-0.389	-0.456
1ubqA	3.11	-0.264	-0.397
3rn3A	2.07	-0.128	-0.224
451cA	2.68	-0.229	-0.406

b)	Λ_E	Λ_e	Λ_{e2}
1iroA	3.67	-0.525	-0.814
1ubqA	2.70	-0.209	-0.329
3rn3A	1.99	-0.103	-0.169
451cA	2.52	-0.189	-0.312

Table 3.3: Lagrange multipliers determined from evolutionary averaged energies. **a)** Lagrange parameters obtained from solving linear equation eq. (3.43) (for coefficients and evolutionary averaged energies for protein chain 3rn3A see Table 3.2). **b)** Lagrange parameters obtained by fitting to correlations for comparison.

The coefficients c_{i0} are not prefactors of Lagrange parameters and represent the mean energies obtained if amino acid substitutions were not correlated, i.e., Q is equal to one, but with the same site-specific probabilities $P_i(a)$ as obtained from the simulation. Table 3.2 lists coefficient for the protein 3rn3. While the coefficients c_{10} and c_{20} are close to $\left\langle \left[U\right] \right\rangle_{\mathrm{evol}}$ and $\left\langle \left[U^2\right] \right\rangle_{\mathrm{evol}}$, respectively, the coefficient c_{00} is considerable larger than $\left\langle E_{\mathrm{nat}} \right\rangle$, suggesting that constraints on the native free energy require strong correlations.

The Lagrange parameters found from the constraints are consistent with the Lagrange parameters obtained from fitting (values are compared in Table 3.3).

3.3.2 Empirical data

The assumption underlying the statistical analysis of protein data is that residue pairs in different proteins which are structurally equivalent have a similar pattern of correlated substitutions, as observed for simulated data. Here, a large non-redundant set of proteins is considered, whose native structures and sequences are taken from the PDB. The correlation pattern is found by grouping structurally equivalent pairs of residues into bins and count amino acid pair frequencies for each bin. This presents an average of the correlation over many different pairs, thereby averaging out correlation patterns specific for individual sites and pairs of sites. Hence, the hope is that the average correlation pattern is generic to all sites as one might expect for folding stability.

Structural equivalence comprises the position of each residue within the contact network as well as the relative position of the two residues. For the latter I introduce a variable c, which is 1 if the pair is in contact and 0 otherwise. Pairs that are closer in sequence than three residues do not contribute to the stability against misfolding and are therefore neglected. For these pairs the variable c is set to -1.

A good way to distinguish residues which is relevant in the context of sequence evolution is the effective connectivity as the site-specific amino acid probability is closely related to the EC. This has not only be shown for simulated data, but also for empirical data in previous studies [24, 64, 66]. Indeed, the present ansatz of grouping pairs and determine pair frequencies is an extension of the aforementioned studies, which groups sites with similar values of the EC and thus determines in this way the site-specific probability as a function of the EC.

In the last Section the correlation measure Q was defined as the ratio of a pair probability of two sites and a product of the site-specific frequencies. However, by grouping pairs of residues the concept of single sites is no longer well defined and a redefinition of the correlation measure is required. A plausible ansatz is to measure the correlation as the ratio of the probability to find a pair of amino acids a and b in the contact class c and the probability to find the pair in all classes except the excluded diagonal,

$$Q = \frac{P(a,b|c,e_1,e_2)}{P(a,b|c\neq -1,e_1,e_2)} \quad . \tag{3.46}$$

What makes this definition favorable is that it presents an unbiased estimator in the sense that if amino acids pairs are randomly assigned to a contact class c the expectation value for Q is one. However, the redefinition requires new marginalization conditions. A condition similar to eqs. (3.20) holds for the new definition for Q,

$$\sum_b Q(a,b|c,e_1,e_2)\,P(a,b|c\neq -1,e_1,e_2) = P(a|c,e_1,e_2) \tag{3.47a}$$

$$\sum_a Q(a,b|c,e_1,e_2)\,P(a,b|c\neq -1,e_1,e_2) = P(b|c,e_1,e_2) \quad . \tag{3.47b}$$

Opposed to the marginalization eq. (3.20) for alignment data, the marginalization here is not satisfied by setting Q to one if the marginalized frequencies $P(a|c\neq -1,e_1,e_2)$ and $P(a|c,e_1,e_2)$ differ. For instance, one has to expect correlations, i.e., $Q\neq 1$ holds if an amino acid participates more often in a contact than expected from randomly assigning pairs to contact classes.

In addition to the marginalization constraints, the correlation measure satisfies the normalization condition,

$$\sum_{c\neq -1} Q(a,b|c,e_1,e_2)\,P(c,e_1,e_2) = P(c\neq -1,e_1,e_2) \quad . \tag{3.48}$$

Consequently, the correlation for non-contacts is completely determined by the correlation for contacts,

$$Q(a,b|c=0,e_1,e_2) - 1 = \big(1 - Q(a,b|c=1,e_1,e_2)\big)\,\frac{P(c=1|e_1,e_2)}{P(c=0|e_1,e_2)} \quad . \tag{3.49}$$

Since more pairs are not in contact than in contact ($P(c=0)\gg P(c=1)$), the correlation Q for non-contacts is expected to be much weaker.

class	probability $P(c)/P(c \neq -1)$
non-contact	0.971
contact	0.029
non-contact EC small-small	0.241
non-contact EC small-large	0.494
non-contact EC large-large	0.237
contact EC small-small	0.004
contact EC small-large	0.008
contact EC large-large	0.017
no direct/no indirect	0.900
no direct/indirect	0.071
direct/no indirect	0.010
direct/indirect	0.019

Table 3.4: Frequencies of pair classes in different binning experiments.

Adapting the theory

As the definition of the correlation measure in eq. (3.46) does not incorporate single sites, it is not strictly compatible with the theory developed in Section 3.2.2. Hence, the theory is modified to fit the correlation measure defined for empirical data.

First, I maximize for each contact class c the relative entropy

$$-\sum_{a,b} P(a,b,c \neq -1, e_1, e_2)\, Q(a,b,c,e_1,e_2) \ln Q(a,b,c,e_1,e_2) \tag{3.50}$$

under the constraints (3.47), which are regulated by the Lagrange multipliers $h(a,c,e_1,e_2)$ and $g(b,c,e_1,e_2)$. I do not consider the normalization (3.48), which would couple the theory for the two contact classes, making it more difficult to solve. If the two EC classes e_1 and e_2 are identical, one cannot distinguish the two amino acids as a matter of principle and every pair quantity is symmetric under swapping of amino acids a and b. Therefore, the number of independent marginalization constraints reduces from fourty to twenty. As a consequence, the two Lagrange multipliers $h(a,c,e,e)$ and $g(a,c,e,e)$ are equal.

Folding stability is now constrained by conditions to the mean and the mean squared energy for each contact class: $\sum_{a,b} P(a,b,c,e_1,e_2)U^i(a,b)$ with $i = 1,2$, which motivates the Lagrange multipliers $\Lambda_i(c)$. The solution to the constrained maximum entropy is now

$$Q_1^{\text{pred}}(a,b,c,e_1,e_2) = h(a,c,e_1,e_2)\, g(b,c,e_1,e_2) \exp\left(\Lambda_1(c)U(a,b) + \Lambda_2(c)U^2(a,b)\right) \quad . \tag{3.51}$$

The Lagrange multipliers $h(a,c,e_1,e_2)$ and $g(b,c,e_1,e_2)$ are determined numerically from the marginalization conditions (3.47), in parallel with the Lagrange parameters $\Lambda_i(c)$, which are found from a fit to Q^{meas} (see below).

Second, similar to the theory developed before, the numeric determination of the parameters $h(a)$ and $g(b)$ can be avoided by a linear theory. However, here the linearization $\ln Q \approx Q - 1$ does not allow for resolving the Lagrange multipliers $h(a,c,e_1,e_2)$ and $g(b,c,e_1,e_2)$ because of

the different marginalization conditions. Instead, the following ansatz is made in analogy with eq. (3.29),

$$Q_2^{\text{pred}}(a,b,c,e_1,e_2) = q(a,b,c,e_1,e_2) + \left(\frac{P(a|e_1)P(b|e_2)}{P(a,b|c\neq -1,e_1,e_2)}\right)\left[\Lambda_1(c)\,u_1(a,b,e_1,e_2)+\Lambda_2(c)\,u_2(a,b,e_1,e_2)\right] \tag{3.52}$$

with the definitions

$$\begin{aligned} u_1(a,b,e_1,e_2) =& U(a,b) - \sum_{a'} P(a'|e_1)U(a',b) - \sum_{b'} P(b'|e_2)U(a,b') \\ &+ \sum_{a',b'} U(a,'\,b')P(a'|e_1)P(b'|e_2) \end{aligned} \tag{3.53a}$$

$$\begin{aligned} u_2(a,b,e_1,e_2) =& U^2(a,b) - \sum_{a'} P(a'|e_1)U^2(a',b) - \sum_{b'} P(b'|e_2)U^2(a,b') \\ &+ \sum_{a',b'} U^2(a,'\,b')P(a'|e_1)P(b'|e_2) \quad . \end{aligned} \tag{3.53b}$$

Apparently, equation (3.29) requires the introduction of the site-specific probability $P(a|e)$, which is defined as the probability of finding a residue within the effective connectivity class e. The prefactor $P(a|e_1)P(b|e_2)/P(a,b|c\neq -1)$ ensures that the terms proportional to the parameters $\Lambda_k(c)$ yield zero if plugged into the marginalization conditions (3.47). In general, the function $q(a,b,c,e_1,e_2)$ cannot be set to one, since this is not a solution of the marginalization conditions (3.47).

The ratio between the product of the marginalized probabilities $P(a|c,e_1,e_2)$ and $P(b|c,e_1,e_2)$ and the pair probability $P(a,b|c\neq -1,e_1,e_2)$ satisfies the marginalization eq. (3.47), but not the normalization condition (3.48), which links $q(a,b,c,e_1,e_2)$ of different contact classes c. However, one can find a $q(a,b,c,e_1,e_2)$ which satisfies both, the marginalization and the normalization conditions,

$$\begin{aligned} q(a,b,c,e_1,e_2) =& 1 + \frac{P(a|c,e_1,e_2)P(b|c,e_1,e_2)}{P(a,b|c\neq -1,e_1,e_2)} \\ &- \sum_{c'\neq -1} \frac{P(c',e_1,e_2)}{P(c\neq -1,e_1,e_2)} \frac{P(a|c',e_1,e_2)P(b|c',e_1,e_2)}{P(a,b|c\neq -1,e_1,e_2)} \quad . \end{aligned} \tag{3.54}$$

Indeed, the function $q(a,b,c,e_1,e_2)$ represents the correlation due to different frequencies of amino acids in contact classes, as discussed above, rather than selection. Therefore, the function q can be regarded as a parameter free theory of correlated mutations. The analogy of q in the exponential theory eq. (3.51) can be found by setting the interaction energy $U(a,b)$ to zero, yielding $Q_1^{\text{pred}}(a,b) = h(a)g(b)$. Even in this simplified theory the Lagrange multipliers have to be determined numerically.

As opposed to the function $q(a,b,c,e_1,e_2)$, the term containing the Lagrange parameters $\Lambda_i(c)$ in eq. (3.52) does not necessarily fulfill the normalization condition. Indeed, since the terms u_1 and u_2 do not depend on the contact class and are in general not linearly dependent, the normalization conditions link the Lagrange parameters for different contact classes,

$$\sum_{c\neq -1} \Lambda_i(c)P(c) = 0, \quad \text{for } i = 1,2 \quad . \tag{3.55}$$

Binning of pairs and fit to data

In the following, I consider the statistical data of proteins taken from the PDB, which contains approximately 80,000 structures comprising 200,000 protein chains. However, protein sequences and structures are not sampled evenly as many sequences and structures are very similar, giving rise to potential sampling bias. Thus, I consider only the rank-one chains of sequence cluster provided by the PDB, which groups sequences with more than 50% sequence identity, thus ensuring that every pair of sequence in my set has less than 50% sequence identity. This set is further filtered to remove inaccurate structures with a high amount of unknown amino acids. The set is not filtered with respect to the method the structure was determined from, in order to ensure sufficient statistics. This has the drawback that rather inaccurate structures that are determined by NMR are found in the set. The remaining 11,720 chains consists of 1,800,758 amino acids and 163,079,832 pairs of amino acids. For each protein chain I count the number $N_p(a,b,c,e_1,e_2)$ of amino acids pairs a, b that can be found in contact class c and in EC class e_1 and e_2. If the two EC classes are identical, the count should be symmetric under swapping the two amino acids. This is achieved by adding the counts to $N_p(a,b,c,e,e)+N_p(b,a,c,e,e)$. Thus, I yield only 210 different counts and therefore the corresponding 210 pairs with $a \leq b$ are considered in the following.

Finally, the counts for all chains p are added and the probability $P(a,b|c,e_1,e_2)$ is computed as

$$P(a,b|c,e_1,e_2) = \frac{\sum_p N_p(a,b,c,e_1,e_2)}{\sum_p \sum_{a',b'} N_p(a',b',c,e_1,e_2)} \quad . \tag{3.56}$$

The EC-specific single site probability is computed accordingly,

$$P(a|e) = \frac{\sum_p N_p(a,e)}{\sum_p \sum_{a'} N_p(a',e)} \quad , \tag{3.57}$$

where $N_p(a,e)$ is the number of amino acids a in the EC class e.

To estimate the statistical error of Q, I assume the counts summed over all proteins to be Poisson distributed, i.e., the standard deviation of the counts is equal to the square root of the counts. Using Gaussian error propagation, I find the error of Q,

$$\begin{aligned}\frac{\Delta Q(a,b,c,e_1,e_2)}{Q(a,b,c,e_1,e_2)} = \Bigg(& \frac{1}{N(a,b,c,e_1,e_2)} + \frac{1}{\sum_{a',b'} N(a',b',c,e_1,e_2)} \\ & + \frac{1}{N(a,b,c\neq -1,e_1,e_2)} + \frac{1}{\sum_{a',b'} N(a',b',c\neq -1,e_1,e_2)} \Bigg)^{1/2} \quad . \end{aligned} \tag{3.58}$$

The statistical error presents only a lower estimate for the true error, as systematic errors, e.g. a sampling bias, might play a role as well. However, these errors are difficult to estimate.

The Lagrange multipliers $\Lambda_i(c)$ are found by minimizing the weighted root mean square deviation (wRMSD) of the prediction,

$$\mathrm{wRMSD}^2 = \frac{1}{n_d} \sum_{a,b} \sum_{e_1,e_2} \frac{(Q_{\mathrm{meas}}(a,b,c,e_1,e_2) - Q_{\mathrm{pred}}(a,b,c,e_1,e_2))^2}{\Delta Q^2_{\mathrm{meas}}(a,b,c,e_1,e_2)} \quad , \tag{3.59}$$

where n_d is the number of data points going into the two sums. Similar to the χ^2_{red} known from statistics, the wRMSD is expected to be in the order of one, assumed that the theory describes the data well (if the proper treatment of degrees of freedoms is ignored for the moment).

In the case of the linearized theory (3.52), the minimization problem is solved by a standard weighted linear fit with $1/\Delta Q^2$ as weights (here, I use the implementation of the GNU GSL library). For the exponential theory (3.51), however, the marginalization conditions require the Lagrange multipliers $h(a)$ and $g(b)$ to change if $\Lambda_i(c)$ changes, which renders the search for the minimal wRMSD difficult. Nevertheless, a simple iterative scheme succeeds: At the beginning the Lagrange parameters $\Lambda_i(c)$ are set to zero and the Lagrange multipliers $h(a)$ and $g(b)$ are determined numerically from marginalization constraints. Here, I choose to minimize the square deviation from the correct marginalization with a conjugate gradient method (implemented by GNU GSL library), which converges quickly. Then, I minimize the wRMSD by a conjugate gradient search of the parameters $\Lambda_i(c)$ while $h(a)$ and $g(b)$ are held constant. The iteration starts now again by determining $h(a)$ and $g(b)$, now in accordance with Λ_i obtained from the last step. The iteration scheme converges within only a few dozen steps.

Besides the wRMSD I introduce a second measure for the quality of the fit, the weighted Pearson correlation coefficient (wCC), which is essentially the Pearson correlation coefficient, but with computing the averages with $1/\Delta Q^2$ as weights,

$$\text{wCC}(Q_{\text{meas}}, Q_{\text{pred}}) = \frac{\left\langle Q_{\text{meas}} Q_{\text{pred}} \right\rangle - \left\langle Q_{\text{meas}} \right\rangle \left\langle Q_{\text{pred}} \right\rangle}{\sqrt{\left\langle Q^2_{\text{meas}} \right\rangle - \left\langle Q_{\text{meas}} \right\rangle^2} \sqrt{\left\langle Q^2_{\text{pred}} \right\rangle - \left\langle Q_{\text{pred}} \right\rangle^2}} \tag{3.60}$$

$$\text{with } \langle x \rangle = \frac{\sum_{a,b} \sum_{e_1,e_2} x(a,b,e_1,e_2)\, 1/\Delta Q^2_{\text{meas}}(a,b,e_1,e_2)}{\sum_{a,b} \sum_{e_1,e_2} 1/\Delta Q^2_{\text{meas}}(a,b,e_1,e_2)} \quad ,$$

thereby giving more weight to data points with a small statistical error.

Two contact classes

In order to keep things simple and attain a small statistical error, the first test is to divide the pairs in only two classes, namely contacts and non-contacts, which were found to be the major determinant of correlation in the previous Section.

Although I have argued that the definition of Q is unbiased if the pairs are randomly assigned to contact classes, it is instructive to test the bias of the estimator with three tests: First, I design random sequences by independently drawing amino acids from a distribution reflecting the amino acid frequencies found in the PDB (see Table A.3 on page 94). I find a Q that indistinguishable from one within the estimated error bars, i.e., the estimator is unbiased. Second, I shuffle amino acid sequences, thereby keeping the amino acid content of each sequence unchanged. Third, each pair of residues is randomly assigned to the contact and the non-contact class, but with keeping the total number of contacts fixed and not changing the amino acid sequence. The last two tests produce similar patterns for $Q(a,b,c)$, which are significantly different from one. Especially, I yield a Q considerably larger than one for cysteine-cysteine pairs in the contact class.

This bias arises, because chains of different lengths are combined when adding the pair counts in (3.56). The fraction of pairs, which form a contact, and the amino acids composition of protein sequences depend on chain length. The number of possible contacts, i.e., the number of

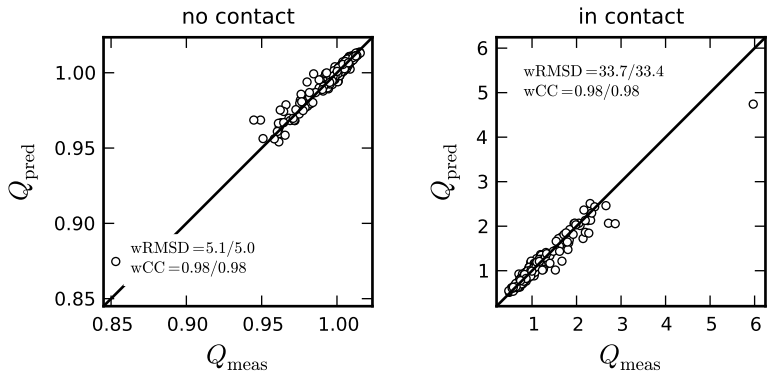

Figure 3.8: Q_{meas} vs. Q_{pred} from fit of linear theory for statistical data to two contact classes.

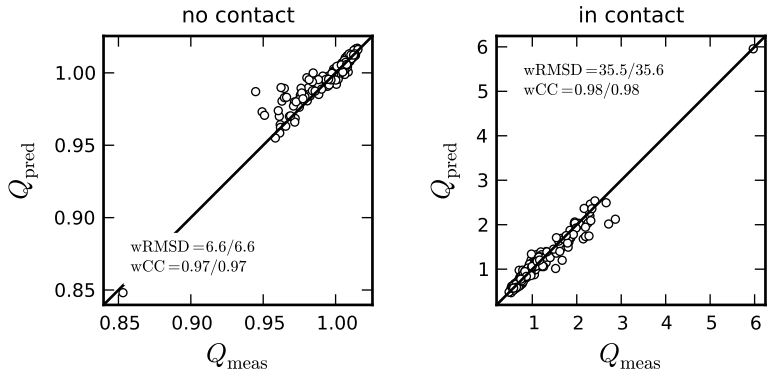

Figure 3.9: Q_{meas} vs. Q_{pred} from fit of exponential theory to statistical data with two contact classes.

pairs, is proportional to the chain length squared and the number of contacts is roughly proportional to chain length. Thus, the fraction of contact pairs with respect to all pairs decreases with chain length. In wild type protein sequences the content of the amino acid cysteine is large in proteins with short chains and decreases with chain length, leveling off at a chain length of approximately 100 residues [32, 67]. Thus, cysteine rich sequences with a relative large fraction of contact pairs and sequences with a low cysteine content and a small fraction of contacts are combined in the computation of Q, yielding a $Q(\text{cys}, \text{cys}) \approx 1.3$ for shuffled sequences and pairs. This bias, however, can be interpreted in a physical meaningful way, as it reflects the positive selection of cysteine-cysteine contacts, i.e., the selection for disulfide bridges in small proteins. Other amino acids do not suffer as much from this bias as cysteine, because their frequency hardly depends on chain length.

The bias can be avoided if the correlation Q is computed for each protein separately and is averaged over all proteins. However, along with this local estimator comes another bias. If Q is computed for each protein, the counts $N_p(a, b, e_1, e_2, c)$ can become zero, resulting in a Q smaller than one even for random sequences. Thus, between the two biases I select the less problematic one.

The correlation of wild type sequences, measured for the two contacts classes, is much larger in scale than the pattern observed for shuffled sequence. The standard deviation of Q is 17 times larger for wild type sequences than for random sequences. Even $Q(\text{Cys}, \text{Cys})$ is approximately six, i.e., significant larger than the value of 1.3 found in the shuffling experiment. Of course,

	linear theory		exp. theory	
	fit to contact	fit to non-contact	fit to contact	fit to non-contact
Λ_1	-1.04	0.0283	-1.374	0.0454
Λ_2	1.66	-0.0607	-0.235	-0.0386
wCC(U=0)	0.86(0.87)	0.87(0.90)	0.86(0.87)	0.81(0.84)
wRMSD(U=0)	85.6(83.5)	13.7(11.6)	84.4(82.6)	16.1(14.5)
wCC(fit)	0.98(0.98)	0.98(0.98)	0.98(0.98)	0.97(0.97)
wRMSD(fit)	33.7(33.4)	5.1 (5.0)	35.5(35.6)	6.6 (6.6)
ΔAIC	$[1.3(1.2)] \cdot 10^6$	$[3.4(2.3)] \cdot 10^4$	$[1.2(1.2)] \cdot 10^6$	$[4.6(3.5)] \cdot 10^4$
ΔBIC	$[1.3(1.2)] \cdot 10^6$	$[3.4(2.3)] \cdot 10^4$	$[1.2(1.2)] \cdot 10^6$	$[4.6(3.5)] \cdot 10^4$

Table 3.5: Fit parameters, weighted RMSD, and weighted correlation coefficient. The data in parenthesis denote the values without the cysteine-cysteine pair. The values for $U = 0$ assess the parameter free theory, $q(a, b, c, e_1, e_2)$ and $h(a, c, e_1, e_2)g(b, c, e_1, e_2)$ respectively, which only account for the marginalization conditions.

such a strong signal is little surprising, as attractive amino acids are often found in contact in the native structure of a protein. Nevertheless, it is interesting to assess whether the theory can reproduce the observed correlation pattern.

Fig. 3.8 and Fig. 3.9 show the results of fitting the linear and exponential theory to contacts and non-contacts. Even though the correlation coefficient wCC is large, the error wRMSD of the fits is much larger than one, therefore, the theories cannot describe the correlation pattern within error bars. Table 3.5 lists the parameters found in the fitting and the wRMSD and wCC for the parameter free and two parameter theories. Surprisingly, the parameter free theories attain a large correlation coefficient and an error, which is only slightly larger than for the two parameter theories. The exception is the cysteine-cysteine pair, for which the prediction of the parameter free theories is not significantly different from other amino acid pairs. Since the parameter free theories only account for the correct marginalization to the probabilities $P(a|c)$, the marginalization has to largely determine the correlation pattern.

This can be understood from the observation that the pair probability $P(a, b|c)$ is rather well approximated by the product of the marginalized probabilities $P(a|c)P(b|c)$ (the correlations coefficient (cc) is $cc = 0.95$ for contacts and $cc = 1.00$ for non-contacts). The probability $P(a, b|c \neq -1)$ is almost perfectly approximated by the product of its marginalized frequencies $P(a|c \neq -1)P(b|c \neq -1)$, and hence the correlation Q for contacts can very well approximated ($cc = 0.86$),

$$Q(a, b|c = 1) \equiv \frac{P(a, b|c = 1)}{P(a, b|c \neq -1)} \approx \frac{P(a|c = 1)}{P(a|c \neq -1)} \frac{P(b|c = 1)}{P(b|c \neq -1)} \quad ,$$

with the exception of cysteine-cysteine contacts (data not shown). The same holds true for non-contacts, where the term $P(a|c = 0)P(b|c = 0)/P(a|c \neq -1)P(b|c \neq -1)$ is a very good approximation of the measured Q ($cc = 0.80$). The ratio $P(a|c = 1)/P(a|c \neq -1)$ can be interpreted as the contact propensity of amino acid a. The contact propensity of an amino acid can be very large (1.6 for tryptophan) and is very well correlated with the hydrophobicity ($cc = 0.89$, data not shown). Thus, the correlations for (non-)contacts is very well approximated by the product of the (non-)contact propensities.

In contrast to contacts, the correlation of non-contacts is rather poorly approximated by the term $P(a|c = 0)P(b|c = 0)/P(a, b|c \neq -1)$, which is part of the $q(a, b, c)$ term. This means,

while the good prediction of the parameter free maximum theory $h(a)g(b)$ is explained by the good approximation of product of (non-)contact propensities, the good prediction quality for non-contacts of the $q(a,b,c)$ term is owed to the fact that the correlation measures for contact and non-contacts are linked by the normalization condition. Indeed the term $P(a|c=1)P(b|c=1)/P(a,b|c\neq -1)$ is largely anti-correlated with $q(a,b,c=0)$, while the corresponding term for $c=0$ is not very well correlated.

Both, the correlation coefficient wCC and the fitting error wRMSD, improve upon fitting the Lagrange parameters $\Lambda_i(c)$. The fitting error decreases to half of the value of the parameter free theories and the correlation coefficients are almost perfect (see Table 3.5). Since the parameter free theory gives already very good results, it is worthwhile to check whether the improvement in prediction justifies the increased complexity of the model. To this end, I compare the two theories by means of the Akaike (*AIC*) and the Bayesian information criteria (*BIC*),

$$AIC = -2\ln L + 2k \quad \text{and} \quad BIC = -2\ln L + 2k\ln n \quad , \tag{3.61}$$

where L is the likelihood of the prediction, n is the number of data points, and k the number of fit parameters, which equals 2 or 0. Assuming Gaussian error statistics, the likelihood is computed from the error for the prediction: $-2\ln L = n\,\mathrm{wRMSD}^2$. Apparently, the *BIC* penalizes models with more free parameters more than the *AIC*. The difference of the information criteria between the parameter free and the two parameter model is positive, indicating that the two parameter model describes the correlation better despite two free parameters (see Table 3.5). Nevertheless, the difference in the information criteria is mostly due to the difference in the log-likelihood, i.e., $n\,\mathrm{wRSMD}^2$. However, the wRMSD is very large because the statistical error of Q has a small estimate. With a larger and more realistic estimate of the error the two parameter model would be less favorable with regard to the information criteria.

The parameter $\Lambda_1(1)$ is large and negative for contacts, in accordance with the expectation that attracting pairs have a larger propensity to be found on contact. For non-contacts, the parameter $\Lambda_1(0)$ is small and negative, and complies with the expectation that non-contacts are rather repulsive. The values of the parameter $\Lambda_2(1)$ obtained from the linear and exponential theory differ largely. Even though the large positive value of $\Lambda_2(1)$ nicely complies with the normalization condition (3.55), it has a rather poor statistical evidence. First, the estimated error of Λ_2 is around five times larger than Λ_2, although in absolute terms it is only around 0.002 due to the small estimated error of Q. Second and more importantly, if I perform a fit of the linear theory, where I set $\Lambda_2(1)$ explicitly to zero, the resulting fitting error is only as large as 42.4(39.9), which is only marginally worse than the fit with $\Lambda_2(1)$. This is true if $\Lambda_2(0)$ is set to zero, where the fit error is only 7.4(5.5). Since the cysteine-cysteine pair is an outlier and $U(\mathrm{cys},\mathrm{cys})$ is large, one might expect that the large value of Λ_2 is caused by the cysteine-cysteine pair. However, if I fit without the cysteine-cysteine pair, I still obtain a large positive value for Λ_2.

Interestingly, the fit parameter $\Lambda_2(1)$ from the exponential theory has to opposite sign compared to the value obtained from the linear theory. Indeed, both for the contact and the non-contact fit, the value is small and negative, indicating that strongly interacting pairs are avoided. Again, the statistical error of the parameter Λ_2 is significantly larger than for Λ_1. For the parameter Λ_1 the values from linear and exponential theory agree, indicating that attractive pairs are preferred in contact and avoided in non-contacts, similar to the results of simulated data.

In summary, the difference between parameter free and two parameter theory captures a signal that is ascribed to selection for folding stability.

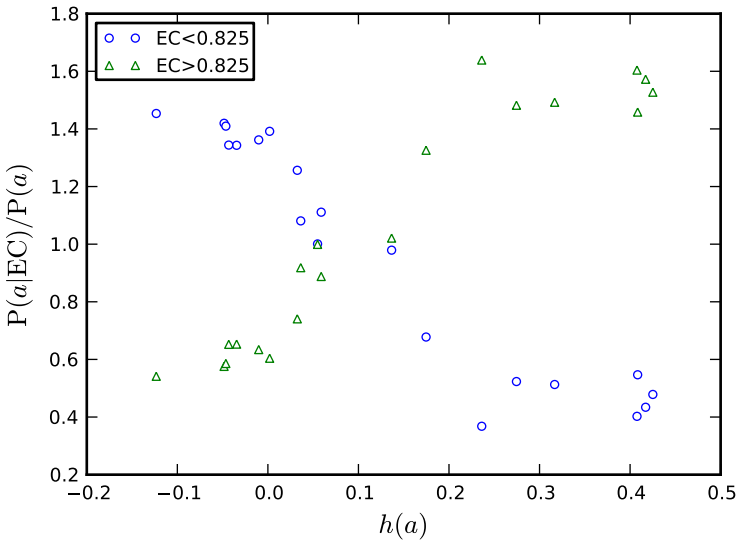

Figure 3.10: Single site frequencies of EC classes. The relative frequency $\mathrm{P}(a|\mathrm{EC})$ of the amino acids a (in-)decreases relative to the background frequency $\mathrm{P}(a)$ for sites with its hydrophobicity at sites with a large(small) EC-value.

EC classes

In the following, residues are distinguished by their EC-value, i.e., by their position in the contact network. To avoid a large statistical error, only two EC classes are considered, where the EC-value is either smaller or larger than the median of the EC-value distribution, which measured value is 0.865. From all combinations of the two EC classes arise three different EC-pair classes: *small-small*, *small-large*, and *large-large*. Each EC-pair class is subdivided into a contact and a non-contact class, yielding six pair classes.

The splitting has only a small effect on the correlation Q of the *small-small* and *large-large* EC classes, which are symmetric under swapping of residues. These are highly correlated to the Q of contacts and non-contacts considered in the last Section. The *small-large* EC classes, which is not symmetric with respect to swapping residues, correlates neither well with the contact nor with the non-contact class.

The contact propensity of a residue depends now on the EC classes, i.e., it is defined as $\mathrm{P}(a|c=1,e_1,e_2)/\mathrm{P}(a|c\neq -1,e_1,e_2)$. As before, the approximation of the correlation by the product of contact propensities is very well for all classes ($cc > 0.8$ without the outlying cysteine-cysteine pair) with the exception of the *large-large* EC classes ($cc \approx 0.7$). Accordingly, the prediction quality of the parameter free theory is lowest for the *large-large* EC classes (Table 3.7).

For the linear fit, the frequency of amino acids is restricted to an EC class $\mathrm{P}(\mathrm{a}, EC)$. Fig. 3.10 shows the ratio of $\mathrm{P}(\mathrm{a}, EC)$ and the background frequency $\mathrm{P}(a)$ versus the hydrophobicity of the amino acids a. In accordance with previous studies, the amino acids with a higher hydrophobicity are found at sites with a large EC-value, while amino acids with a small hydrophobicity are found at a site with a small EC-value.

The fit is now performed for the two contact classes, where the sum error of the respective EC classes is minimized. The wRMSD of the *small-large* EC class is generally in the order of one

and therefore much smaller than for the other two classes, since there are 400 instead of 210 residues pairs, which reduces the counts in the bins and hence increases the estimated error of Q. This shows that a further splitting of classes would increase the statistical noise, such that the fit cannot be applied properly. Accordingly, the fit cannot reduce the wRMSD for the *small-large* EC class by an significant amount and improves the correlation coefficient wCC only marginally, as opposed to the symmetric EC classes. In fact, the fit works well even for the *large-large* EC class, which was poorly predicted by the parameter free theory, and raises the wCC to a high value comparable to the other EC classes as well as it reduces the wRMSD significantly.

The fit parameters are similar to the values obtained in the last Section for just two contact classes. This is not very surprising, since the fit parameters are mainly determined by the symmetric EC classes, which have a correlation pattern similar to the two contact classes from the last Section. Nevertheless, the results show that the theory works well even if the data becomes more detailed.

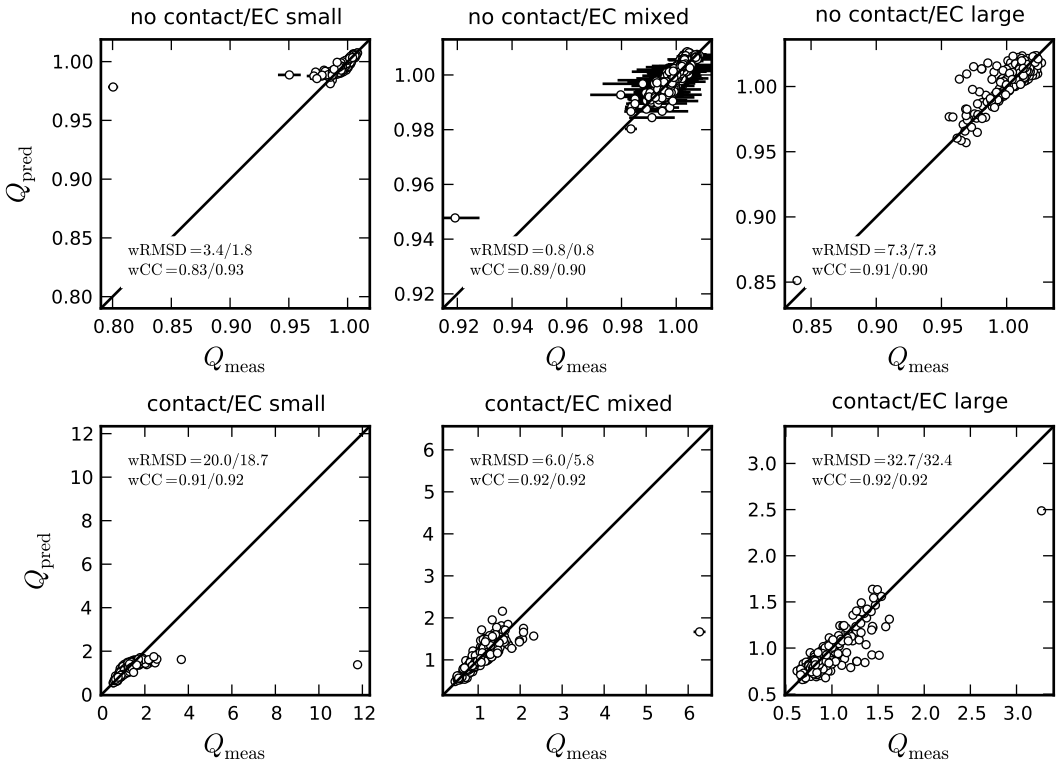

Figure 3.11: Q_{meas} vs. Q_{pred} from fit of linear theory for statistical data to two contact and two EC classes.

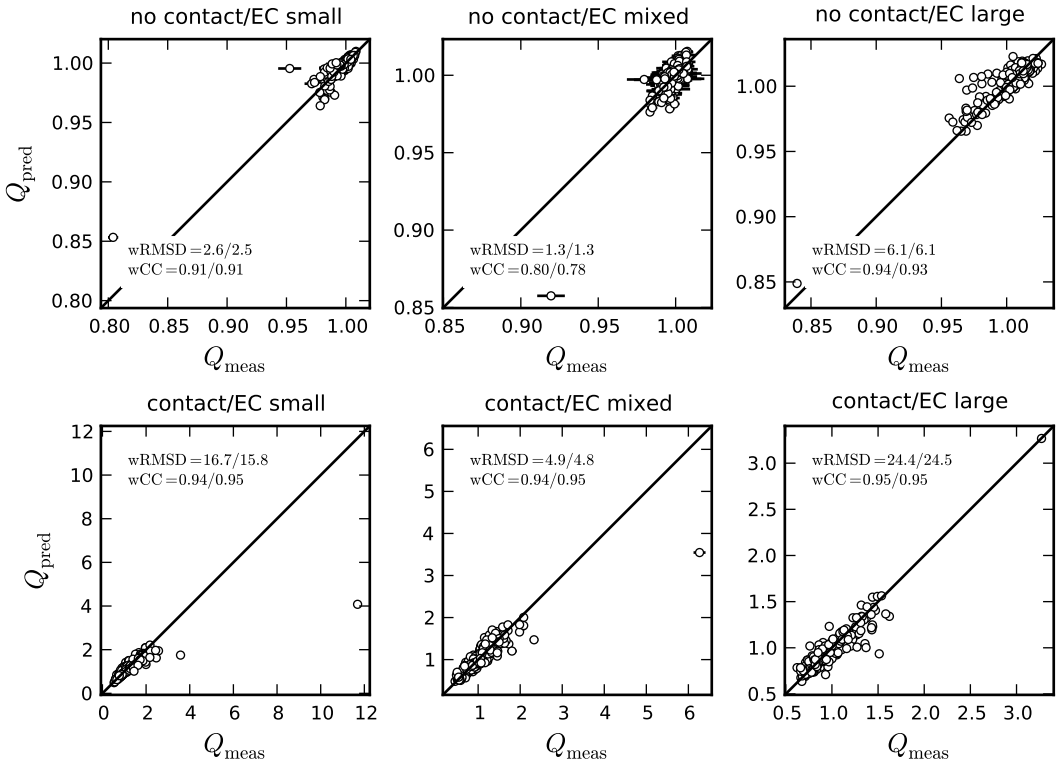

Figure 3.12: Q_{meas} vs. Q_{pred} from fit of exponential theory for statistical data to two contact and two EC classes.

	linear theory		exp. theory	
	fit to contact	fit to non-contact	fit to contact	fit to non-contact
Λ_1	-0.647	0.018	-1.05	0.0336
Λ_2	0.039	-0.032	-0.175	-0.0447
wRMSD($U = 0$)	28.9(27.8)	6.2(5.1)	28.6(27.7)	6.4(5.5)
wRMSD(fit)	19.9(19.4)	4.1(3.8)	15.3(15.1)	3.5(3.5)

Table 3.6: Fit parameters to data binned into two contact class and three EC-pair classes.

class	wRMSD(U=0)	wCC(U=0)	wRMSD(fit)	wCC(fit)
linear theory - fit to contact				
contact/EC small	26.9(26.0)	0.84(0.85)	20.0(18.7)	0.91(0.92)
contact/EC mixed	7.0(6.8)	0.88(0.89)	6.0(5.8)	0.92(0.92)
contact/EC large	49.5(47.6)	0.79(0.80)	32.7(32.4)	0.92(0.92)
average	28.9(27.8)		19.9(19.4)	
ΔAIC = 3.6e+05(3.3e+05)	ΔBIC = 3.6e+05(3.3e+05)			
linear theory - fit to non-contact				
no contact/EC small	4.3(2.9)	0.70(0.82)	3.4(1.8)	0.83(0.93)
no contact/EC mixed	0.8(0.7)	0.83(0.86)	0.8(0.8)	0.89(0.90)
no contact/EC large	11.4(9.7)	0.77(0.82)	7.3(7.3)	0.91(0.90)
average	6.2(5.1)		4.1(3.8)	
ΔAIC = 1.8e+04(9.6e+03)	ΔBIC = 1.8e+04(9.6e+03)			
exp. theory - fit to contact				
contact/EC small	26.7(25.8)	0.84(0.85)	16.7(15.8)	0.94(0.95)
contact/EC mixed	7.0 (6.8)	0.88(0.89)	4.9 (4.8)	0.94(0.95)
contact/EC large	48.9(47.3)	0.79(0.80)	24.4(24.5)	0.95(0.95)
average	28.6(27.7)		15.3(15.1)	
ΔAIC = 4.8e+05(4.4e+05)	ΔBIC = 4.8e+05(4.4e+05)			
exp. theory - fit to non-contact				
no contact/EC small	4.3 (3.0)	0.69(0.80)	2.6(2.5)	0.91(0.91)
no contact/EC mixed	0.8 (0.7)	0.83(0.87)	1.3(1.3)	0.80(0.78)
no contact/EC large	11.8(10.3)	0.75(0.79)	6.1(6.1)	0.94(0.93)
average	6.4(5.5)		3.5(3.5)	
ΔAIC = 2.4e+04(1.5e+04)	ΔBIC = 2.4e+04(1.5e+04)			

Table 3.7: List of wRMSD and wCC for two EC classes.

	linear theory			
	direct/no indirect	direct/indirect	no direct/no indirect	no direct/indirect
Λ_1	-0.849	-0.858	0.0284	-0.0637
Λ_2	-0.104	1.736	-0.0736	0.108
wRMSD($U=0$)	35.6(35.3)	67.2(64.4)	16.5(13.5)	19.9(19.5)
wCC($U=0$)	0.77(0.78)	0.94(0.94)	0.98(0.99)	1.00(1.00)
wRMSD(fit)	23.5(23.5)	32.3(31.0)	8.1(7.8)	18.1(18.1)
wCC(fit)	0.91(0.91)	0.99(0.99)	1.00(1.00)	1.00(1.00)
ΔAIC	$[1.5(1.5)]\cdot 10^5$	$[7.3(6.7)]\cdot 10^5$	$[4.3(2.6)]\cdot 10^4$	$[1.4(1.1)]\cdot 10^4$
ΔBIC	$[1.5(1.5)]\cdot 10^5$	$[7.3(6.7)]\cdot 10^5$	$[4.3(2.6)]\cdot 10^4$	$[1.4(1.1)]\cdot 10^4$
	exp. theory			
	direct/no indirect	direct/indirect	no direct/no indirect	no direct/indirect
Λ_1	-0.953	-1.300	0.0813	-0.0851
Λ_2	-0.294	-0.214	-0.0542	-0.0901
wRMSD($U=0$)	34.5(34.3)	63.5(61.2)	27.3(26.3)	17.7(17.7)
wCC($U=0$)	0.79(0.79)	0.95(0.95)	0.95(0.95)	1.00(1.00)
wRMSD(fit)	20.0(20.0)	30.5(30.5)	13.2(13.0)	16.9(16.9)
wCC(fit)	0.93(0.93)	0.99(0.99)	0.99(0.99)	1.00(1.00)
ΔAIC	$[1.7(1.6)]\cdot 10^5$	$[6.5(5.9)]\cdot 10^5$	$[1.2(1.1)]\cdot 10^5$	$[5.8(5.9)]\cdot 10^3$
ΔBIC	$[1.7(1.6)]\cdot 10^5$	$[6.5(5.9)]\cdot 10^5$	$[1.2(1.1)]\cdot 10^5$	$[5.8(5.9)]\cdot 10^3$

Table 3.8: Fit results of four contact classes distinguishing pairs with contacts and indirect contacts.

Indirect contacts

The results from simulation suggest the correlation of pairs not in contact to be largely influenced by the distance in the contact network. In the following, I distinguish between pairs of residues that either have an indirect contact or do not. That is, the *in contact* and *no contact* class are further split into two contacts classes, which distinguish between pairs which have an indirect contact or not. Even though there is no motivation to split the *in contact* class from the simulated data, the splitting is retained for reasons of symmetry. Thus, four contact classes are defined: *direct/no indirect*, *direct/indirect*, *no direct/no indirect*, *no direct/indirect*.

The correlation Q of the *direct/no indirect* contact class is a special case, since the correlation for this contact class does not correlate well with neither the *in contact* class or *no contact* class. It is rather uncommon for pairs in direct contact not to find a third residue that establishes also an indirect contact. In fact, there are almost twice as many contacts with an indirect contact than without (see Table 3.4). Furthermore, an indirect contact seems to increase the correlation between residues in contact, as the *direct/indirect* class has larger values of Q than the *direct/no indirect* class.

The *no direct/no indirect* class correlates well with the *no contact* class and the *direct/indirect* contact class correlates well with the *in contact* class, meaning that these new classes do not differ significantly from the old ones.

Only 7.1% of all residue pairs belong to the *no direct/indirect* class (see Table 3.4). Interestingly, the correlation for indirect contacts correlates better with the contact class than with the no contact class and has significantly larger values for Q than the *no direct/no indirect* class, which can be interpreted as the manifestation of indirect correlations. The correlation for indi-

rect contacts is already very well predicted by the parameter free theories and yields a perfect wCC, while the wRMSD is still as large as 17-19 due to the small estimated error. The fit cannot reduce the wRMSD significantly, probably due to the almost perfect prediction of the parameter free theory. Consequently, the ability of the theory to describe indirect contacts and the fit parameters for the *no direct/indirect* class cannot be interpreted reliably.

With exception of the *direct/no indirect* contact class all classes are astonishingly well predicted by the parameter free theories. All classes except the *no direct/indirect* contact class improve significantly upon fitting (cf. Table 3.8, Fig. 3.13, and Fig. 3.14). With exception of the parameter Λ_2 for the *direct/no indirect* class found from the linear fit, the parameters agree very well parameters found for the *contact* and *non-contact* classes, respectively.

However, the actual aim of this section, namely to test the theory for indirect contacts, was undermined by the fact that the *no direct/indirect* class is very well predicted by the parameter free theory.

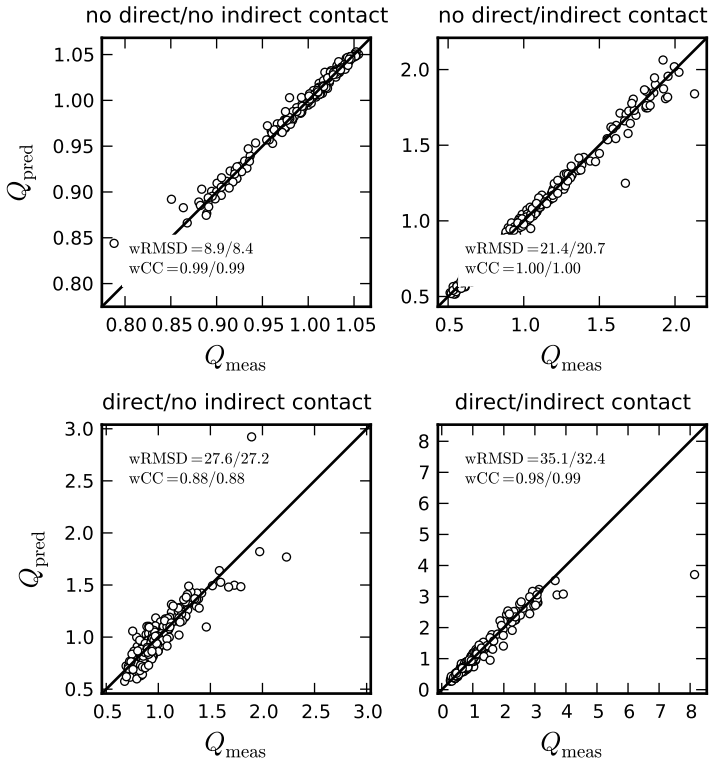

Figure 3.13: Linear fit to indirect contact classes.

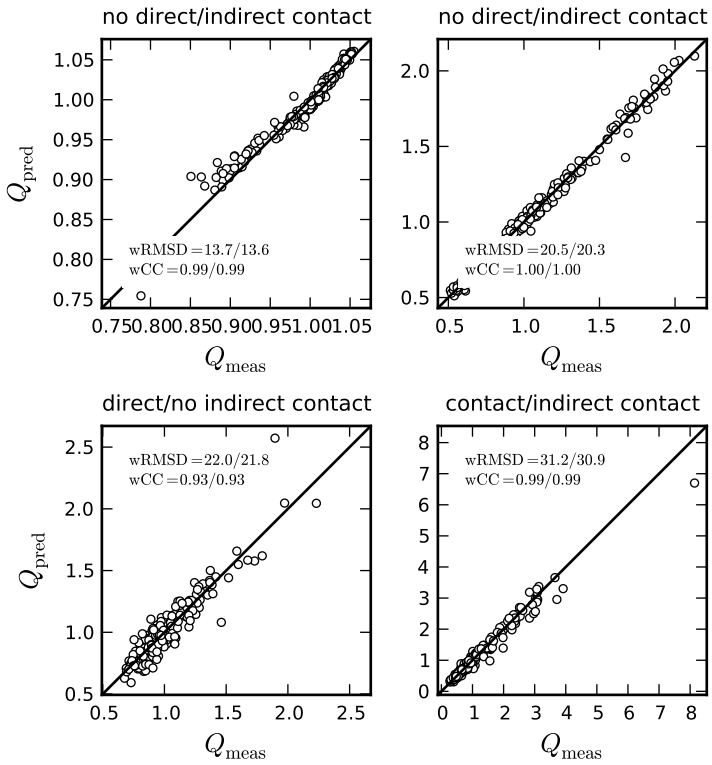

Figure 3.14: Exponential fit to indirect contact classes.

3.4 Discussion

A model for correlated amino acids substitutions is introduced, whose basic principle is the maximization of the sequence entropy. The sequence entropy is expanded in a cluster expansion up to the pairwise term. Thus, the model produces a prediction of the correlation that depends on the direct interaction of two residues in the native and non-native states.

Correlated substitutions, which are measured from simulations, allow a detailed test of the model. Positive design gives rise to strong correlations of residues involved in a native contact, which are very well predicted by the model. The ability to predict non-native contacts varies considerably for two reasons: First, the strong correlation of native contacts causes indirect correlations, which have a major contribution to the correlation of residues indirectly connected in the contact network. It is possible to improve the model with an *ad hoc* term, which accounts for the indirect correlations. However, a proper treatment of indirect correlations would demand more effort. A possible way is to consider a cluster expansion that involves three residues, which should be able to consider indirect correlations due to indirect contacts. However, such an analytical computation is cumbersome and does not account for indirect correlations that are due to an indirect contact via two or more contacts. An alternative is to compute the indirect correlations numerically from marginalizing of probability of the entire sequence, eq. (3.11), similar to methods used in the work of Weigt *et al.* [53]. However, this requires a large computational effort and would give little insight.

Indirect correlations should not contribute to the correlation of pairs of residues that are distant in the contact network. This assumption is supported by the observation that the average MI levels off at large distances in the contact network. Thus, it is reasonable to assume that the correlation at these distances is due to the direct interaction of the residues in the misfolded ensemble. However, the prediction quality of the model for these pairs is rather poor. This cannot be explained by the fact that the fit of the model is determined by other pairs, which have a greater weight in the fit, since, if the fit is performed for each pair of residues separately, the quality of the fit for large distant pairs does not improve. This brings us to the second reason why the model does not fit contacts well. The constraint in the model and the constraint used in the simulations are not identical. In the model imposes constraints on the mean and standard deviation of the free energy misfolded of the misfolded ensemble, whereas in the simulations the energy gap α is constrained. Such a possible mismatch could be clarified by changing the constraints in the simulations. Nevertheless, a good prediction of non-native contacts is not very important as the correlation of native contacts yields information about the native state. Besides, the correlation of non-native contacts is rather weak.

The model has three parameters, which are found by fitting to the observed correlation or from the constraints, which produce compatible values. The values for the Lagrange parameters are in agreement with our expectations: Native contacts show a strong positive correlation of attractive and strong negative correlation of repulsive contacts, i.e., the Lagrange parameter Λ_E is positive and large. The conditions against misfolding, imposed by the REM, determines the correlation of non-native contacts. To increase the mean free energy of the misfolded energy, the attractive amino acids have to be suppressed, i.e., the parameter Λ_e is negative. The free energy of the misfolded ensemble is increased if the standard deviation of the free energy of misfolded conformations is reduced, because then less misfolded structures with negative free energies

occur. Accordingly, amino acids with a strong interaction are found to be anti-correlated, i.e. the parameters Λ_{e2} is negative.

Of course, it is important to validate the model for empirical data. The correlation was measured as an average over many pairs of residues by binning residues from many proteins by their position in the contact network. After a minor adaption due to the changed definition of the correlation, the model helps to reproduce the amino acid pair frequency in different structural classes. However, it is found that the theory could not be tested for indirect contacts, as their correlation is almost perfectly predicted by the trivial part of theory.

It should be noted that the test for empirical data is somewhat self referential as the interaction energies are fitted to natural proteins and hence depend on the contact propensity of amino acid pairs. Nevertheless, the good quality of the prediction indicates the model for correlated mutations and the interaction model are consistent. Of greater interest is evolutionary data for individual proteins, which allows to investigate correlated substitutions for individual residues, which will be discussed in the next Chapter.

4 Conclusion and Outlook

The thermodynamic stability of protein structures is an important selection pressure in protein sequence evolution. In Chapter 2 it was shown that the stability against misfolding is selected in wild type sequences, where contact frequency and contact correlations arising in the misfolded ensemble play a significant role. Before, misfolding stability was estimated by the REM, which neglects contact correlations. Furthermore, it was shown that the consideration of the third cumulant improves the approximation of the free energy, as it accounts for deviations of the free energy distribution from a Gaussian approximation, which is the central assumption of the REM. Thus, the model of protein folding can be made more realistic by considering the cumulant expansion of the misfolded free energy up to the third order, whose cumulants are determined by statistical properties of the contacts in misfolded structures.

Such an improved model permits to investigate negative design at a greater detail. The native state breaks the symmetry of pairs and hence it is reasonable to assume that some pairs are under stronger selection pressure due to negative design than others. As a next step, it would be interesting to predict, which residues and pairs of residues are under particularly strong selection of negative design.

Negative design has consequences for the correlated substitutions of amino acids, as a strong selection of pairs of residues due to negative design should result in a strong correlation. The model for correlated mutations in its current version, however, incorporates the misfolded ensemble using the REM. Thus, the model could be made more realistic by constraining the free energy difference $\Delta G = E_{\text{nat}} - G_{\text{misfold}}$, where G_{misfold} is estimated from the cumulant expansion and respects contact correlations as discussed above.

Simulations of sequence evolution that constrain the detailed free energy present a prefect tool to investigate correlated mutations and to assess the predictions of the model. In addition, simulations may help to test prediction schemes for pairs of residues under strong selection from negative design.

Nevertheless, the strongest correlations are found for native contacts, which were predicted well by the model. This property of the model might be exploited for the prediction of contacts from correlated mutations, which presents are inversion of the model. The model allows to compute the likelihood of the observed pair frequency $P_{ij}(a, b)$ under the assumption that i and j are in contact, which could serve as a score to distinguish contacts from non-contacts. Such a prediction scheme could be tested using data generated from simulations.

For a bioinformatical application, however, this scheme would have to be applied to evolutionary data of natural proteins, contained in MSA. However, MSA data suffers from sampling biases, that is, some clusters of similar sequences occur more often than others. This bias can relatively easily be compensated by reweighting the sequences. More problematic are correlated mutations that arise from the common ancestry of the sequences rather than selection forces. While some authors claim that reweighting sequences is sufficient to reduce this bias [51], other authors claim that more intricate methods for the cleaning of the data are necessary [49, 50, 63], which they have applied successfully . However, these methods where tested only for correlation measures that assess the overall correlation of two sites. The advantage of the model introduced in this thesis is that it considers amino acid pair specific information, given for instance by $Q_{ij}(a, b)$. Therefore, it would be necessary to adapt the data preprocessing methods

to amino acid pair specific data. Again, the performance of this cleaning could be tested on data produced by simulations.

Since the model of correlated mutations is based on pairwise interactions of residues, it cannot predict indirect correlations, making a further improvement of the model necessary. In fact, indirect correlations have a large contribution and hence would have to be respected in a contact prediction scheme. The model of Haw *et al.* [51, 53], which is similar to ours, infers direct interaction parameters between residues to disentangle direct from indirect correlations, yielding an enhanced quality of the contact prediction. The model presented in thesis is, however, more detailed, as it describes the correlation of individual amino acids pairs. Thus, the model could be used to improve the scheme of Hwa *et al.* by combining the direct information and the likelihood of a contact into one score.

In summary, in this thesis it was shown how methods of statistical physics can help to investigate questions of protein folding and protein sequence evolution, some of which even have bioinformatical applications.

A Appendix

first nucleotide	second nucleotide: U		second nucleotide: C		second nucleotide: A		second nucleotide: G	
U	UUU	F	UCU	S	UAU	Y	UGU	C
	UUC	F	UCC	S	UAC	Y	UGC	C
	UUA	L	UCA	S	UAA	*	UGA	*
	UUG	L	UCG	S	UAG	*	UGG	W
C	CUU	L	CCU	P	CAU	H	CGU	R
	CUC	L	CCC	P	CAC	H	CGC	R
	CUA	L	CCA	P	CAA	Q	CGA	R
	CUG	L	CCG	P	CAG	Q	CGG	R
A	AUU	I	ACU	T	AAU	N	AGU	S
	AUC	I	ACC	T	AAC	N	AGC	S
	AUA	I	ACA	T	AAA	K	AGA	R
	AUG	M	ACG	T	AAG	K	AGG	R
G	GUU	V	GCU	A	GAU	D	GGU	G
	GUC	V	GCC	A	GAC	D	GGC	G
	GUA	V	GCA	A	GAA	E	GGA	G
	GUG	V	GCG	A	GAG	E	GGG	G

Table A.1: The standard genetic code (* $\simeq$ stop codon)

A	4	L	6	R	6	F	2
E	2	G	4	T	4	Y	2
Q	2	K	2	P	4	C	2
D	2	S	6	I	3	W	1
N	2	V	4	M	1	H	2

Table A.2: Codon degeneracies, i.e. the number of codons that encode one amino acid, according to the standard genetic code.

amino acid	frequency	amino acid	frequency
ALA	0.0814	ARG	0.0532
GLU	0.0704	THR	0.0530
GLN	0.0371	PRO	0.0432
ASP	0.0561	ILE	0.0612
ASN	0.0397	MET	0.0166
LEU	0.0996	PHE	0.0415
GLY	0.0686	TYR	0.0347
LYS	0.0595	CYS	0.0142
SER	0.0581	TRP	0.0133
VAL	0.0748	HIS	0.0237

Table A.3: Background frequency of amino acids observed in a non redundant subset of the PDB.

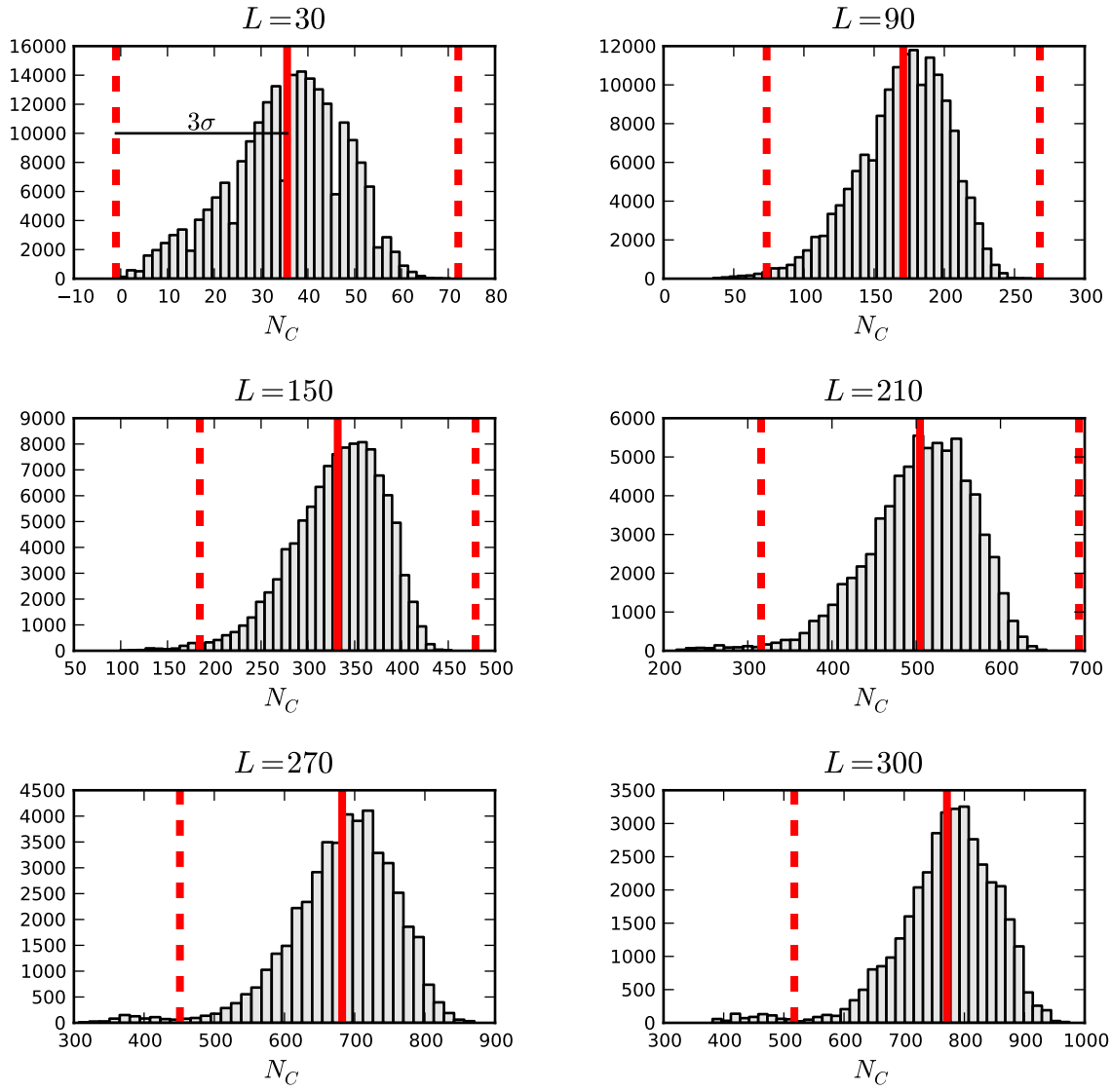

Figure A.1: Distribution of number of contacts and filtering of non-compact structures for different query lengths L. All substructures generated by threading, whose number of contacts is more than three standard deviations apart from the mean, are considered non-compact and removed from the set.

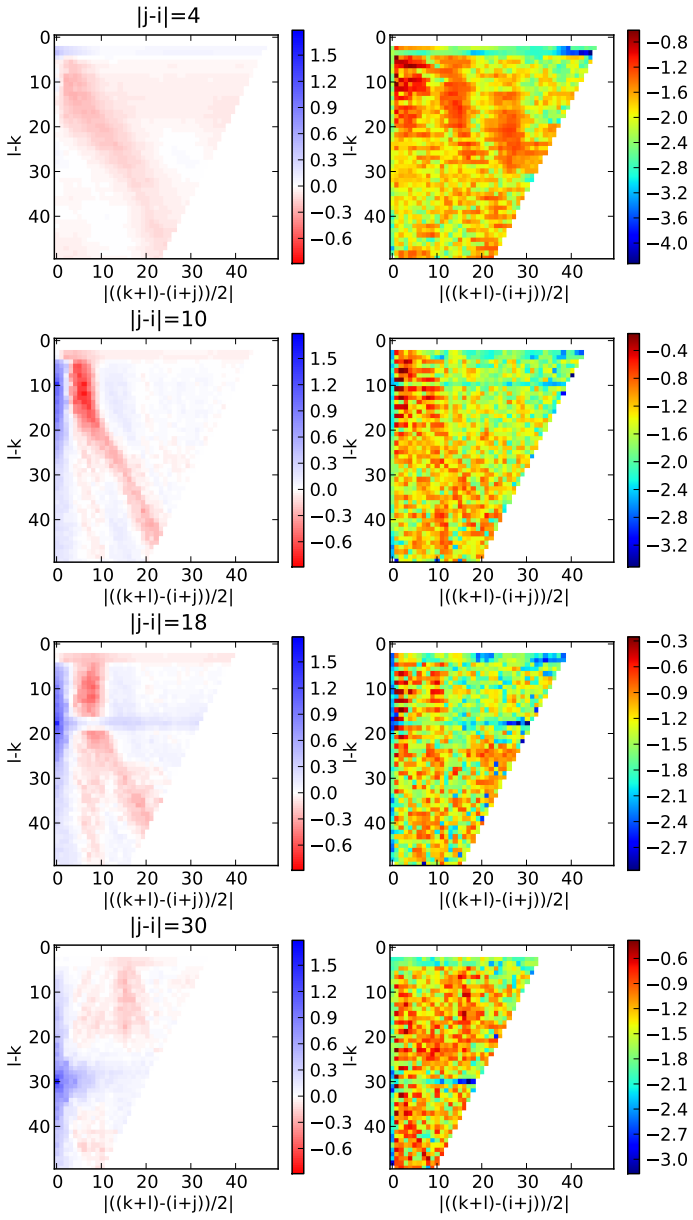

Figure A.2: Contact correlation from threading binned into homogeneous indices. Left column: mean value of contact correlation. Right column: Logarithm of the ration of standard deviation and the mean of the bin. Obviously, the relative standard deviation is very small in most bins. Therefore, the homogeneous approximation is justified.

Bibliography

[1] C. B. Anfinsen, Principles that govern the folding of protein chains. *Science*, **181**(4096), (1973), 223–230.

[2] C. M. Dobson, Protein folding and misfolding. *Nature*, **426**(6968), (2003), 884–890.

[3] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, and H. Ploegh, *Molecular Cell Biology*. W. H. Freeman and Company, sixth edition edition, 2008.

[4] D. Graur and W.-H. Li, *Fundamentals of molecular evolution*. Sinauer Associates, second edition edition, 2000.

[5] J. S. Richardson, D. C. Richardson, N. B. Tweedy, K. M. Gernert, T. P. Quinn, M. H. Hecht, B. W. Erickson, Y. Yan, R. D. McClain, and M. E. Donlan, Looking at proteins: representations, folding, packing, and design. biophysical society national lecture, 1992. *Biophys J*, **63**(5), (1992), 1185–1209.

[6] W. DeLano, The pymol molecular graphics system, 2002.

[7] K. A. Dill, Dominant forces in protein folding. *Biochemistry*, **29**(31), (1990), 7133–7155.

[8] M. Vendruscolo, E. Kussell, and E. Domany, Recovery of protein structure from contact maps. *Fold Des*, **2**(5), (1997), 295–306.

[9] M. Vendruscolo, R. Najmanovich, and E. Domany, Protein folding in contact map space. *Phys. Rev. Lett.*, **82**, (1999), 656–659.

[10] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, The protein data bank. *Nucl. Acids Res.*, **28**(1), (2000), 235–242. ISSN 0305-1048.

[11] U. Bastolla, J. Farwer, E. W. Knapp, and M. Vendruscolo, How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins*, **44**(2), (2001), 79–96.

[12] U. Bastolla, M. Porto, and A. R. Ortíz, Local interactions in protein folding determined through an inverse folding model. *Proteins*, **71**(1), (2008), 278–299.

[13] M. A. DePristo, D. M. Weinreich, and D. L. Hartl, Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet*, **6**(9), (2005), 678–687.

[14] R. A. Goldstein, The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins*, **79**(5), (2011), 1396–1407.

[15] D. M. Taverna and R. A. Goldstein, Why are proteins marginally stable? *Proteins*, **46**(1), (2002), 105–109.

[16] L. Holm and C. Sander, Mapping the protein universe. *Science*, **273**(5275), (1996), 595–603.

[17] K. Illergård, D. H. Ardell, and A. Elofsson, Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins*, **77**(3), (2009), 499–508.

[18] B. Rost, Protein structures sustain evolutionary drift. *Fold Des*, **2**(3), (1997), S19–S24.

[19] E. Zuckerkandl, Evolutionary processes and evolutionary noise at the molecular level. i. functional density in proteins. *J Mol Evol*, **7**(3), (1976), 167–183.

[20] E. A. Franzosa and Y. Xia, Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol*, **26**(10), (2009), 2387–2395.

[21] A. Tóth-Petróczy and D. S. Tawfik, Slow protein evolutionary rates are dictated by surface-core association. *Proc Natl Acad Sci U S A*, **108**(27), (2011), 11151–11156.

[22] R. Sasidharan and C. Chothia, The selection of acceptable protein mutations. *Proc Natl Acad Sci U S A*, **104**(24), (2007), 10080–10085.

[23] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins*, **58**(1), (2005), 22–30.

[24] U. Bastolla, A. R. Ortíz, M. Porto, and F. Teichert, Effective connectivity profile: a structural representation that evidences the relationship between protein structures and sequences. *Proteins*, **73**(4), (2008), 872–888.

[25] F. Teichert, U. Bastolla, and M. Porto, Sabertooth: protein structural alignment based on a vectorial structure representation. *BMC Bioinformatics*, **8**, (2007), 425.

[26] F. Teichert, J. Minning, U. Bastolla, and M. Porto, High quality protein sequence alignment by combining structural profile prediction and profile alignment using saber-tooth. *BMC Bioinformatics*, **11**, (2010), 251.

[27] K. Wolff, M. Vendruscolo, and M. Porto, Efficient identification of near-native conformations in ab initio protein structure prediction using structural profiles. *Proteins*, **78**(2), (2010), 249–258.

[28] S. Maisnier-Patin, O. G. Berg, L. Liljas, and D. I. Andersson, Compensatory adaptation to the deleterious effect of antibiotic resistance in salmonella typhimurium. *Mol Microbiol*, **46**(2), (2002), 355–366.

[29] A. Poon, B. H. Davis, and L. Chao, The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. *Genetics*, **170**(3), (2005), 1323–1332.

[30] O. Noivirt-Brik, R. Unger, and A. Horovitz, Analysing the origin of long-range interactions in proteins using lattice models. *BMC Struct Biol*, **9**, (2009), 4.

[31] I. N. Berezovsky, K. B. Zeldovich, and E. I. Shakhnovich, Positive and negative design in stability and thermal adaptation of natural proteins. *PLoS Comput Biol*, **3**(3), (2007), e52.

[32] U. Bastolla and L. Demetrius, Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng Des Sel*, **18**(9), (2005), 405–415.

[33] Y. K. Mok, C. M. Kay, L. E. Kay, and J. Forman-Kay, Noe data demonstrating a compact unfolded state for an sh3 domain under non-denaturing conditions. *J Mol Biol*, **289**(3), (1999), 619–638.

[34] B. Shan, D. Eliezer, and D. P. Raleigh, The unfolded state of the c-terminal domain of the ribosomal protein l9 contains both native and non-native structure. *Biochemistry*, **48**(22), (2009), 4707–4719.

[35] D. Shortle, The denatured state (the other half of the folding equation) and its role in protein stability. *The FASEB Journal*, **10**(1), (1996), 27–34.

[36] J. U. Bowie, R. Lüthy, and D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**(5016), (1991), 164–170.

[37] M. P. Morrissey and E. I. Shakhnovich, Design of proteins with selected thermal properties. *Fold Des*, **1**(5), (1996), 391–405.

[38] B. Derrida, Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, **24**(5), (1981), 2613–2626.

[39] J. D. Bryngelson and P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci U S A*, **84**(21), (1987), 7524–7528.

[40] E. I. Shakhnovich and A. M. Gutin, Formation of unique structure in polypeptide chains. theoretical investigation with the aid of a replica approach. *Biophys Chem*, **34**(3), (1989), 187–199.

[41] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics*, **21**(3), (1995), 167–195. ISSN 1097-0134.

[42] Pande, Grosberg, Joerg, and Tanaka, Is heteropolymer freezing well described by the random energy model? *Phys Rev Lett*, **76**(21), (1996), 3987–3990.

[43] I. N. Berezovsky, A. Y. Grosberg, and E. N. Trifonov, Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett*, **466**(2-3), (2000), 283–286.

[44] M. M. G. Krishna and S. W. Englander, The n-terminal to c-terminal motif in protein folding and function. *Proc Natl Acad Sci U S A*, **102**(4), (2005), 1053–1058.

[45] Deutsch and Kurosky. New algorithm for protein design, *Phys Rev Lett*, **76**(2), (1996), 323–326.

[46] Seno, Vendruscolo, Maritan, and Banavar, Optimal protein design procedure. *Phys Rev Lett*, **77**(9), (1996), 1901–1904.

[47] W. Jin, O. Kambara, H. Sasakawa, A. Tamura, and S. Takada, De novo design of foldable proteins with smooth folding funnel: automated negative design and experimental verification. *Structure*, **11**(5), (2003), 581–590.

[48] U. Göbel, S. Chris, R. Schneider, and A. Valenciax, Correlated mutations and residue contacts in proteins. *Proteins*, **18**, (1994), 309–317.

[49] S. D. Dunn, L. M. Wahl, and G. B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24(3)**, (2008), 333–340.

[50] L. Burger and E. van Nimwegen, Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology*, **6**.

[51] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, **108**(49), (2011), E1293–E1301.

[52] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3d structure computed from evolutionary sequence variation. *PLoS One*, **6**(12), (2011), e28766.

[53] M. Weigt, R. A. White, H. Szurmantc, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing. *PNAS*, **106(1)**, (2009), 67–72.

[54] O. Noivirt-Brik, A. Horovitz, and R. Unger, Trade-off between positive and negative design of protein stability: From lattice models to real proteins. *PLoS Computational Biology*, **5(12)**.

[55] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan, Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structual Biology*, **10(1)**, (2003), 59–69.

[56] J. Baussand and A. Carbone, A combinatorial approach to detect coevolved amino acid networks in protein families of variable divergence. *PLoS Computational Biology*, **9**.

[57] S. W. Lockless and R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**(5438), (1999), 295–299.

[58] B.-C. Lee, K. Park, and K. Dongsup, Analysis of the residue–residue coevolution network and the functionally important residues in proteins. *Proteins*, **72**, (2008), 863–872.

[59] G. Sella and A. E. Hirsh, The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A*, **102**(27), (2005), 9541–9546.

[60] H. A. Bethe, Statistical theory of superlattices. *Proc. Royal Soc. Lon.*, **150**, (1935), 552–575.

[61] J. Hijmans and J. de Boer, An approximation method for order-disorder problems. *physica*, **21**, (1955), 471–484.

[62] J. S. Yedidia, W. T. Freeman, and Y. Weiss, Bethe free energy kikuchi approximations and belief propagation algorithms, http://www.merl.com/papers/docs/TR2001-16.pdf, 2001.

[63] O. Noivirt, M. Eisenstein, and A. Horovitz, Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Engineering, Design & Selection*, **18(5)**, (2005), 247–253.

[64] M. Porto, H. E. Roman, M. Vendruscolo, and U. Bastolla, Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. *Mol Biol Evol*, **22**(3), (2005), 630–638.

[65] F. R. Kschischang and B. J. Frey, Factor graphs and the sum-product algorithm. *IEEE TRANSACTIONS ON INFORMATION THEORY*, **47(2)**, (2001), 498–519.

[66] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the protein data bank. *BMC Evol Biol*, **6**, (2006), 43.

[67] S. White, Amino acid preferences of small proteins. implications for protein stability and evolution. *J. Mol. Biol.*, **4**, (1992), 991–995.

Resumé

Personal details

Name:	Jonas Minning
Birth:	10.01.1983 in Bendorf am Rhein

Academic education

2009/02 - present	PhD studies at the Technische Unsiversität of Darmstadt
2007/12 - 2008/12	Diploma thesis entitled “Proteinstrukturvergleich und Proteinsequenz-/ strukturzuordnung”
2005/09 - 2006/06	Erasmus stay at the University of Bath (Great Britain)
2003/04 - 2009/02	Diploma studies in physics at the Technische Universität Darmstadt (Germany)

Civilian service

2002/05 - 2003/02	Herz-Jesu-Krankenhaus Dernbach, Hol- und Bringdienst

Research and Teaching experience

2007-2011	Teaching assistant in theoretical physics and theoretical sold state physics
2009/02	Research stay with the group of Ugo Bastolla at the Centro di Biología Molecular “Severo Ochoa” in Madrid
2009/11	Research stay with the group of Professor Michele Vendruscolo at the University of Cambridge (Great Britain)

Publications

Teichert F., Minning J., Bastolla U., und Porto M.,
High quality protein sequence alignment by combining structural profile prediction and profile alignment using SABERTOOTH, *BMC Bioinformatics*, **11**, (2010), 251

Acknowledgment

I gratefully acknowledged the excellent advice of my supervisor Markus Porto, which was provided even over the large distance between Darmstadt and Cologne. I want to thank Ugo Bastolla for invaluable discussions over the internet. He contributed many important ideas to the two projects. In particular, he developed the theory of correlated mutations.

I also want to thank Barbara Drossel for integrating me into her group and the group members of the AG Drossel for providing a social environment, making me feel less isolated as single member of the AG Porto in Darmstadt. In particular, I want to thank Christopher "Imperator des Spieleabends/Mifo" Priester and Katrin Wolff for a collegial office community, the former I owe special thanks for the many books.

I am indebted to Esther Minning, Eva Ackermann, Korinna Allhoff, Lotta Heckmann, Christof Wolf, and Michael Harrach for proof reading the manuscript.

I also want to thank Christoph Schmitt for his contributions to the work of correlated mutations in the early stage of the project, as well as Florian Teichert and Andreas Buhr for writing libraries that allow to handle biological data and simulate protein sequence evolution. They made my work easier.

Last but not least, I want to thank my parents, who always supported me and made my academic education possible.

Erklärung zur Dissertation

Hiermit versichere ich, die vorliegende Dissertation ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den June 6, 2012

(Jonas Minning)